Care Records and Healthcare Processes: Adding Context to Clinical Codes

Lara Chammas
University of Oxford
lara.chammas@st-annes.ox.ac.uk

Owen P Dwyer
University of Oxford
owen.dwyer@gtc.ox.ac.uk

Emanuel Sallinger
TU Wien
sallinger@dbai.tuwien.ac.at

Jim Davies
University of Oxford
jim.davies@cs.ox.ac.uk

Eva JA Morris
University of Oxford
eva.morris@ndph.ox.ac.uk

Abstract

Process mining techniques are being used to explore healthcare processes based upon information recorded about individual patients. In most cases, this information consists of clinical codes and dates: codes used to classify care events; dates indicating when these events occurred. These codes will not, in general, form part of the contemporaneous care record used by clinicians. At the same time, that record contains other, more detailed information about the care delivered. This paper explains how the provenance of coded information can affect its interpretation and how information from a care record can be used to stratify patient populations and provide context for process mining. The proposed methodology is illustrated through application to real-world data in an area of particular concern: the treatment and care of patients with colon cancer.

Keywords: process mining, care records, healthcare processes, care pathways.

1. Introduction

The widespread adoption of electronic care records (or electronic patient records) has the potential to revolutionise healthcare delivery and accelerate health research (Goldacre & Morley, 2022). A greater degree of automation is needed to realise this potential: in particular, in the analysis of patient journeys to determine whether care is being delivered as intended, to make better use of resources, and to understand how patterns of care affect outcomes (Foley & Vale, 2022).

Process mining techniques are being used to address exactly this need. De Roock and Martin (2022) present an extensive review and make five recommendations for the field: that the work should be driven by specific research questions or needs; that domain experts should be involved; that there should be increased emphasis upon the reporting of data preparation techniques; that there is a need for more research translating findings into actions; and that process mining analysis should consider key performance indicators.

Each of these recommendations reflects the importance of context. A specific question provides a basis for selecting different kinds of source data and for deciding upon an appropriate classification of events. Domain expertise is needed to ensure that not only the question but also the semantics of the data—its provenance and interpretation—is properly understood. An account of semantics and preparation is needed to determine the applicability and generalisability of the results to the healthcare sector.

Another review, Rojas et al. (2016), draws a useful distinction between mining treatment processes, within or across clinical settings, and mining organisational processes, such as the capacity of a clinic; it draws a distinction also between data from clinical systems and data from administrative systems. The need for contextualisation is raised also in Batista and Solanas (2018): in particular, in addressing the heterogeneity of treatment pathways for the same disease.

Events or episodes that have been ‘coded’—that is, associated with a code from a standard terminology—help to reduce this heterogeneity, providing a classification of diseases or treatments. They serve as an excellent basis for further abstraction, as shown by Cremerius et al. (2022), Kurniati et al. (2018), and Remy et al. (2020). However, the interpretation of these codes may depend upon the purpose of the coding activity, and more detailed information—beyond the codes—may be needed to address specific questions.

URI: https://hdl.handle.net/10125/106829
978-0-9981331-7-1

Epidemiology is focused on measuring the distribution and determinants—the "who, where, when"—of disease and other health-related events (dos Santos Silva, 1999). At its most basic it describes the rates of occurrence of a disease, noting differences in groups based on who they are, where they live, and when they lived. It can also determine risk factors of a disease and measure their effect on health outcomes such as life expectancy. Epidemiological methods and research findings can help determine areas of context to explore within healthcare processes.

In this paper, we explain how we can make greater use of context in healthcare process mining. In Section 2, we explore the content of the care record and its relationship to other sources of data within healthcare organisations. In Section 3, we outline an approach that combines process mining techniques with epidemiological research methods, using context from the care record to stratify populations—facilitating discovery, compliance checking, and improvement. Section 4 shows how this approach can be applied to the analysis of treatment processes, addressing a specific research question using real-world data. The paper ends with a discussion, including suggestions for future work.

2. Care records

A care record is a record maintained by healthcare organisations for the purposes of providing care to an individual patient. This is quite distinct from a personal health record maintained by a patient or carer (Lear et al., 2022). The data that it contains may be messy and fragmented, created at different times by different people using different systems. As Goldacre and Morley (2022) observes, it is "historically and by design" an *aide memoire*: a practical record to help manage care.

2.1. Clinical coding

Clinical terminologies are used to save time and facilitate data re-use (NHS Digital, 2021). Instead of writing 'diabetes mellitus', we may enter an ICD-10 code of *E10* or a SNOMED-CT code of *73211009*. Each of these terminologies can be used to provide more detailed information: ICD-10 code *E10* and its ten subdivisions refer only to Type 1 diabetes, with another four top-level diabetes codes to consider.

Our interpretation of the resulting codes will depend crucially upon the context and purpose of the coding process (Nouraei et al., 2016). Coding is undertaken mostly for the purposes of reimbursement or planning: explaining and justifying the consumption of resources, usually for financial reasons. It may also undertaken for the purposes of epidemiology and public health, or for the continuity of care—conveying summary information to other clinicians in an unambiguous form.

The coding process will often involve the resolution of some uncertainty or ambiguity. There may also be decisions needed as to whether and how to include more detailed information. The selection of a code may depend upon purpose and circumstance, as characterised above; it may also be influenced by the technology available (Capita, 2014).

Moreover, the extent to which the coding is informed by clinical expertise may vary between two extremes (Nouraei et al., 2016). At one extreme, a code is assigned by a coder with no medical training and no access to the clinical team, simply on the basis of case notes and coding guidelines. At the other, a code is agreed and assigned by a team of clinicians when they decide upon a treatment plan for the patient in question.

An audit of coding within the UK National Health Service in 2014 (Capita, 2014) revealed significant variation in coding quality: the mean error rate was 8.8% for primary diagnostic codes, and 6.7% for primary procedure codes. The same audit observed that although all clinically-relevant comorbidities should be coded, coders find it difficult to determine whether a given condition is clinically relevant to the primary diagnosis; as a result, many codes are omitted.

For example, there will be patients with diabetes for whom there is no corresponding diagnostic code in their care record, even if their clinical team is aware of their condition and they are being treated for it (Anwar et al., 2011). The clinical team doesn't rely upon the code to convey the information or to determine treatment. For the same reason, there will also be patients coded as Type 1 (*E10*) who are in fact Type 2 (*E11*).

As an example of how these issues may impact our analyses, suppose that we wish to explore the treatment of patients with ketoacidosis, a problem with the body's acid-base balance associated with diabetes, alcoholism, and starvation. The ICD-10 code *E10.1* denotes ketoacidosis in Type 1 diabetics in the absence of a coma, and *E10.0* denotes a diabetic coma, with or without ketoacidosis. From these codes alone, we are unable to determine whether some patients have this condition or not.

The forthcoming ICD-11 will address this particular issue, but a more fundamental problem remains: the codes may not be present, accurate, or consistent. A clinical diagnosis will be based upon a combination of blood test results (Kilpatrick et al., 2022). These results, not the code, will determine the treatment that follows. If our aim is to include ketoacidosis in our analysis of medical treatment processes, in the sense of Lenz and Reichert (2007) and Rojas et al. (2016), then codes alone may not be enough.

2.2. Additional, contextual information

A care record will contain a wide range of documents: notes, forms, reports, letters, observations, results, prescriptions, discharge summaries and more. The design of these documents will, in general, be proprietary: each system supplier will have a different schema and different datatypes, for the representation of the same data (Lenz & Reichert, 2007).

Some degree of standardisation and interoperability is essential for care delivery. To date, industry-driven efforts have focussed upon the development of standards for messaging between systems, such as HL7 FHIR (Fast Healthcare Interoperability Resources): providing access to certain classes of data in a standard format, while preserving the underlying proprietary representations (Lenz & Reichert, 2007).

These standards are enough to support the use of additional, contextual information in process mining. We may extract data such as patient demographics, laboratory test results, and radiology reports, in the same format, from a wide range of clinical systems, regardless of supplier. The provenance and quality of the data may remain an issue, but the data will be available.

There is also the prospect of access to complete care records in a standard form, allowing a more detailed characterisation of healthcare events than coding can provide. Healthcare providers and government agencies are driving the development of 'vendor-neutral' or 'portable' care records, with the intention of mandating their use. This will facilitate the inclusion of *indirect* events in process mining.

This term comes from the ISO 13940 standard, also known as ContSys: a system of concepts to support the continuity of care (International Organization for Standardization, 2015; Oughtibridge, 2019). In ContSys, a *direct* event represents an interaction between a patient and a healthcare professional.

An *indirect* event, or 'indirect healthcare activity period', involves one or more healthcare professionals completing a healthcare activity without the patient being present. An indirect event could be a multidisciplinary team (MDT) meeting to decide upon a treatment plan for a patient with cancer, or the preparation of a prescription. The improved availability of information regarding indirect events will allow the incorporation of patient-specific plans in process mining.

This is already an area of interest for those working on the development and application of vendor-neutral care records. For example, Iglesias et al. (2022) explains how planning information may be used in the exploration of clinical processes based upon care records in the vendor-neutral openEHR standard.

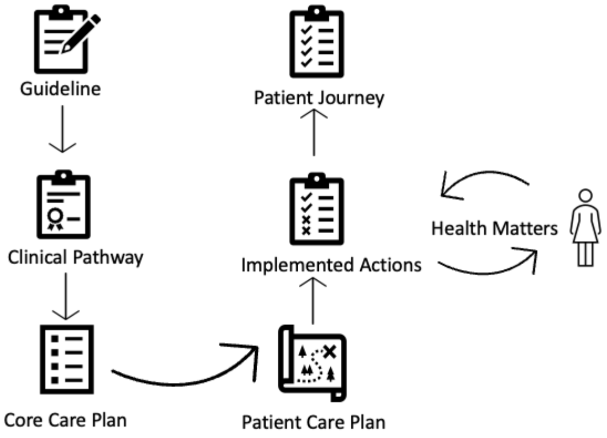

Figure 1. Pathways, plans, and actions

The ContSys standard provides a comprehensive list of terminology related to the continuity of care (Oughtibridge, 2019). There are five terms of particular relevance to process mining (see Figure 1):

clinical guideline a set of statements that assist healthcare professionals in making decisions about which activities to perform for a specific health issue; these are generic and do not concern an actual patient.

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

core care plan a refinement of the clinical pathway, a narrowing of options reflecting the approach taken by a particular organisation; sets out treatment intentions for a group of patients. The corresponding set of traces will be a subset of those for the clinical pathway.

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

patient journey a sequence of healthcare activities for an individual patient; consists of direct events drawn from the care record. The corresponding trace should be one of those allowed for in the patient care plan.

The indirect events in a care record provide information regarding care pathways and plans. For example, a patient care plan may specify a series of future chemotherapy sessions. The direct events describe the actual patient journey. For example, the sessions that were actually delivered.

Existing applications of process mining in healthcare have focused mainly upon patient journeys, as characterised by clinical coding of diagnoses and procedures. The incorporation of indirect events, drawn from care records, will provide an additional basis for exploring processes and assessing compliance, between the level of the care plan and the implemented actions.

3. Method for integrating context

To show how additional information from care records can be used in process mining, we will build upon the method outlined by Cremerius et al. (2022), extending it with aspects of epidemiological research methodology (Chew, 2019; dos Santos Silva, 1999).

The original method outlined by Cremerius et al. (2022) starts with stating a goal, followed by defining patient cohorts and case notation, selection of case attributes, event types and their attributes, and concludes with enriching event attributes. We propose to extend this method with additional steps from clinical research, together with some alignment of terminology, and an increased emphasis upon iteration:

1. define the research question
2. conduct a literature review
3. define the patient population and subcohorts
4. identify the events of interest
5. generate event logs and mine for processes
6. compare results to what is expected
7. investigate deviations
8. present findings to domain experts and iterate

Defining a research question In this approach, questions should concern a particular patient cohort with a particular health issue. This provides for closer alignment with clinical guidelines and care plans, which are developed on the same basis. This facilitates conformance checking, and leads to process models that clinicians can interpret qualitatively. It may also suggest case notation for the event log: some form of case notation will be present in care plans that refer to activities across multiple settings.

Pijnenborg et al. (2021) center their research question on the issue of palliative care treatments for oesophageal and stomach cancer patients. This allowed them to determine common practices in palliative care, link them to survival outcomes, and provide evidence to inform the development of clinical guidelines.

Literature Review In epidemiology, literature reviews involve determining common treatments for a disease, factors that impact disease outcomes or treatment choice, and discovering what are the national and international guidelines for identification, diagnosis, or treatment (Chew, 2019). This will provide a basis for the assessment of our findings, indicate contextual information that may be relevant, and suggest how to define population subcohorts. These reviews are conducted using databases such as PubMed and MedLine.

Patient cohorts Existing approaches to reducing model complexity (Kaymak et al., 2012) are focussed upon event logs and traces: filtering or grouping events, or clustering traces to form trace variant groups for separate analysis (Aspland et al., 2021; Munoz-Gama et al., 2022; Remy et al., 2020).

Contextual information from the care record provides another means of reducing complexity: by stratifying the dataset into cohorts of patients with similar characteristics, who are more likely to have similar journeys; essentially determining case variants before event logs are even generated.

Suitable data points for stratification can be determined during the literature review phase; common ones in epidemiology include age, gender, socio-economic background, and the presence of comorbidites (dos Santos Silva, 1999). Some of these data points may correspond to the occurrence of previous events.

Baker et al. (2017) provide an excellent example of cohort definition, based upon cancer type, chemotherapy drug, and the intent of the chemotherapy treatment. The work also involves the inclusion of health states—additional contextual information—as events in the model: for example, whether a patient's white blood cell count was too low for chemotherapy. The resulting discovery phase is clear and concise.

Events of interest Clinical guidelines and care plans provide a core list of events that should occur in patient journeys. This may prove sufficient for population-level analysis, which can then focus upon the most and least frequent journeys, the average timing between events, and the number of activities involved. Individual-level analysis may require additional events representing comorbidities or health states preceding or following the event log.

To gain an adequate understanding of a particular issue, both forms of analysis may be required. We may need multiple perspectives, and additional context, to understand the significance of deviations from a patient care plan, a core care plan, or a clinical pathway. These will allow us to determine what we should expect to see, and to find areas of deviation for further investigation.

4. Example

As an example of how we may usefully incorporate domain expertise and additional care records data in process mining, we will consider a specific research question of interest: understanding the patient journeys associated with colon cancer treatment at a major English regional hospital.

The dataset used contains anonymous information on the treatment of 2,458 patients with an ICD-10 code of C18 (colon cancer) in their care record. The data covered a period of eight years, from 2012 to 2020, and included information on inpatient and outpatient hospital activities coded using the OPCS-4 terminology (NHS Digital, 2021).

All attendances by the colon cancer patient population at the hospital within this time period are included in the dataset, whether they were related to the colon cancer or not, providing a rich history of events to support our investigation.

Research question and design We set out to determine the different treatment pathways for colon cancer patients in the hospital. We conducted a literature review of the treatments for colon cancer, discovered the factors that have been shown to impact life expectancy, reviewed the relevant clinical guidelines for treatment selection, and reviewed the national clinical pathway.

Our patient population is those with the ICD-10 code (C18) for colon cancer in their care record, filtering for patients with a form of metastatic treatment or a previous cancer diagnosis. As we did not have access to national cancer registry data, we have no information as to whether these diagnoses were confirmed by subsequent national case review; some of these codes may represent a suspected diagnosis for the procedure.

Patient cohort subgroups were defined based on patient age, Charlson score (a measure of comorbidity), and the Indices of Multiple Deprivation (a relative measure of socio-economic deprivation for a small area surrounding the patient's home address). This was based on a literature review suggesting that patient age, comorbidity, and socio-economic status are factors that impact cancer treatment choice and life expectancy (Syriopoulou et al., 2019; Taylor et al., 2021).

We identified four treatment types of interest: major resection, minor resection, chemotherapy, and radiotherapy. The mapping from OPCS codes to our classification of treatments was informed by extensive discussion with colon cancer clinicians and epidemiologists and is aligned with the COloRECTal cancer Repository (CORECT-R) data coding table generated by the UK Colorectal Intelligence Hub, and can be found on the project's webpage ("CORECT-R", 2023).

For initial event log generation, only the four treatment types were included as events. An additional event, cancer incidence, was derived by identifying the earliest date for which a C18 code was recorded. Where this incidence date coincides with a treatment date, the incidence is considered to have occurred first.

Directly follows graphs (DFGs) were generated from the event log for each of the patient cohort subgroups. After identifying pathway anomalies in the graphs, we re-examined the care record to find context that could form the basis of an explanation.

Population level view Figure 2 shows the DFG for the whole cohort of patients. As this is a population-level view, we can draw a comparison with the Cancer Research UK (CRUK) analysis of treatment rates across the country (Cancer Research UK, 2015). In the CRUK analysis, 63% of patients receive a surgical intervention (minor or major resection), 31% of patients receive chemotherapy, 3% receive radiotherapy, either as a single treatment or in combination, and 40% of patients receive no treatment at all.

Our DFG shows that the overall distribution of interventions within the hospital reflects these national averages. However, this high-level view of journeys does not tell us whether the distribution of treatments for particular subgroups of patients is as expected. To explore this, we first stratify the patient cohort by age.

Age Figure 3 shows the DFG for patients aged 50 to 59 at the time of incidence, whilst Figure 4 shows the DFG for those 80 and over. These show that only 26% of patients aged 50 to 59 received no treatment, whereas the figure for those aged 80 or over is 52%.

Furthermore, those aged 80 or over who were treated received only a single form of treatment, rather than a combination. These figures show also that the usage of chemotherapy varies between age groups, with 40% of patients aged 50 to 59 receiving chemotherapy but only 6% of patients aged 80 or over.

This can be explained clinically and through patient preference. The range of treatment options narrows with age as the potential benefit in terms of increased life expectancy is reduced and the cost in terms of side effects is increased due to frailty and comorbidity.

Chemotherapy is recommended only for Stage III or high risk Stage II cancers; older patients with higher stage cancers will be less able to tolerate chemotherapy, and the increase in life expectancy will be minimal.

Younger patients will often prioritise treatment options based upon increased life expectancy; older patients will often prioritise options based upon quality of life (Shrestha et al., 2019). This is reflected in core care plans and patient care plans.

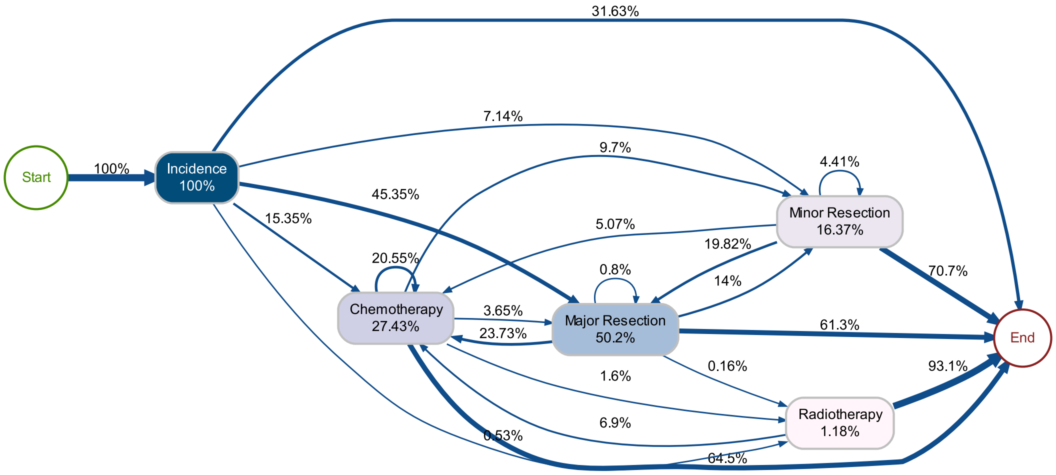

Figure 2. Treatment pathways for all patients with an ICD-10 code of C18

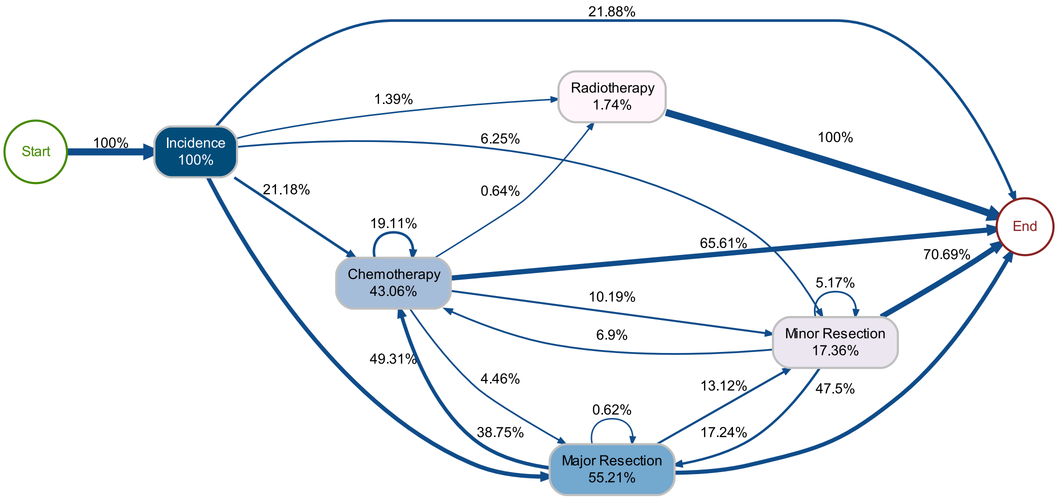

Figure 3. Treatment pathways for all patients with an ICD-10 code of C18 aged 50 to 59

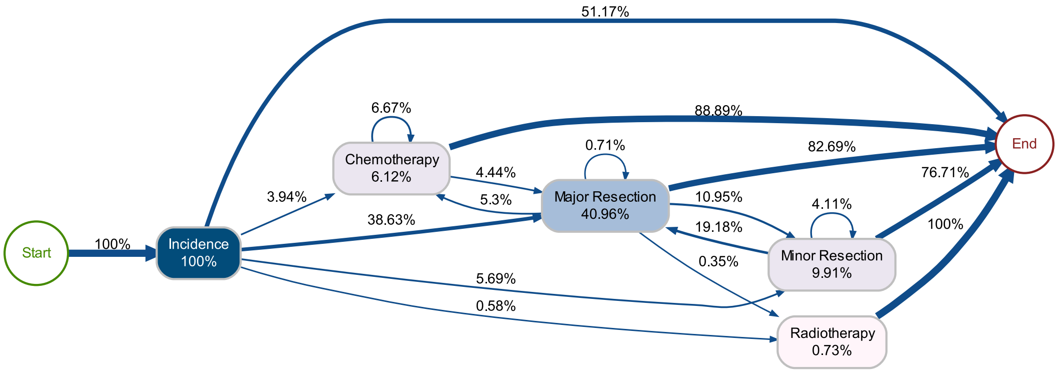

Figure 4. Treatment pathways for all patients with an ICD-10 code of C18 aged 80 and over

Charlson Score We then stratified the patient population using a common measure of comorbidity, the Charlson score, calculated using the *comorbidity* R package (Gasparini, 2018). This score is a summary statistic that takes into account the presence of 17 different diseases associated with mortality risk. It is used in epidemiology to adjust survival models to make research comparable across patient populations. A higher score indicates the presence of more comorbidities, and therefore a higher risk of early mortality. It can also be used as a measure of fragility (Charlson et al., 1987).

When we stratify the patient cohort by Charlson score, we find an interesting pattern: 82% of patients with a Charlson score of 0 (indicating no comorbidities) have no treatment events in their pathway, as seen in Figure 5. This is initially surprising, as increased health is usually associated with the prioritisation of cancer treatments to increase length of life, resulting in more options being consumed (Shrestha et al., 2019).

When we investigate beyond the treatment events of interest and consider other information in the care record, we can see that most of these patients received diagnostic endoscopic investigations of the colon on the basis of "suspicion of malignancy". The C18 code was used to indicate the malignancy in question, and did not represent a confirmed diagnosis. Again, this illustrates the importance of context—in this case, other events in the care record—in supporting analysis.

Socio-economic factors While we would hope to see variations in care on the basis of patient age or fragility, variations associated with socio-economic background are less welcome—and the subject of considerable concern, effort, and investment across public health systems (Syriopoulou et al., 2019).

We stratified our patient population using the English indices of multiple deprivation (IMD), which ranks localities based upon income, employment, education and skills, health and disability, crime, barriers to housing and services, and the living environment.

Of those patients living in the most deprived areas (the first quintile), 57% did not receive any treatment, compared to only 33% in the least deprived areas (the fifth quintile). The major resection rates were 38% and 50% respectively, as shown in Figure 6 and Figure 7.

Time to major resection In England, national guidance recommends that patients with cancer should receive treatment within 62 days of being referred for investigation, on the basis of suspected cancer, by a general practitioner. We considered a subcohort of patients who received a major resection as their first treatment, finding that 45 patients out of 612 (7%) did not receive treatment within the 62 day target.

We then isolated all events between the incidence date and the date of resection for these patients. We found that 35% of these patients received another diagnosis which required treatment, such as anaemia requiring a transfusion, during that time. For another 30%, there were consistent diagnostic imaging or endoscopic events between incidence and resection, indicating a 'wait and watch' approach to the cancer.

These are clinically valid reasons for missing the target, and these applied in the majority of the cases identified. Again, this serves to highlight the importance of considering data points and mining for events beyond the initially-identified clinical codes of interest.

5. Discussion

The challenges of working with real-world health data are well documented: see, for example, Gatta et al. (2018), Martin et al. (2020), Munoz-Gama et al. (2022), and Syed et al. (2023). These include incompleteness, inconsistency, and different levels of granularity. However, with the exception of Fox et al. (2018), where 'issues due to source' are mentioned, there has been relatively little consideration of the context in which clinical codes appear.

In this paper, we have argued that codes alone are not enough to support our analyses and that additional, contextual information is required. We have explored the prospects of obtaining this information from care records and care plans, and explained the potential value of 'indirect events' recorded in patient care plans. We have extended an existing approach to incorporate this information, together with aspects of epidemiological practice, and presented an example involving the use of real-world hospital data.

The extended approach, and the example investigation, were informed by two challenges set out in Munoz-Gama et al. (2022): "looking at the process through the patient eyes" and "considering contextual information when conducting process mining analyses in healthcare". They were informed also by De Roock and Martin (2022), which emphasises the need to "closely incorporate domain experts": we used clinical specialists to help characterise events of interest and interpret the graphs that we obtained.

By considering additional information in the care record, we were able to determine reasons for the observed variations in patient journeys. For example, for the majority of patients whose treatment came later than the national guidelines, we were able to establish that the reason for delay was clinically valid: a competing diagnosis resulting in increased frailty, or a care management decision based upon diagnostic information.

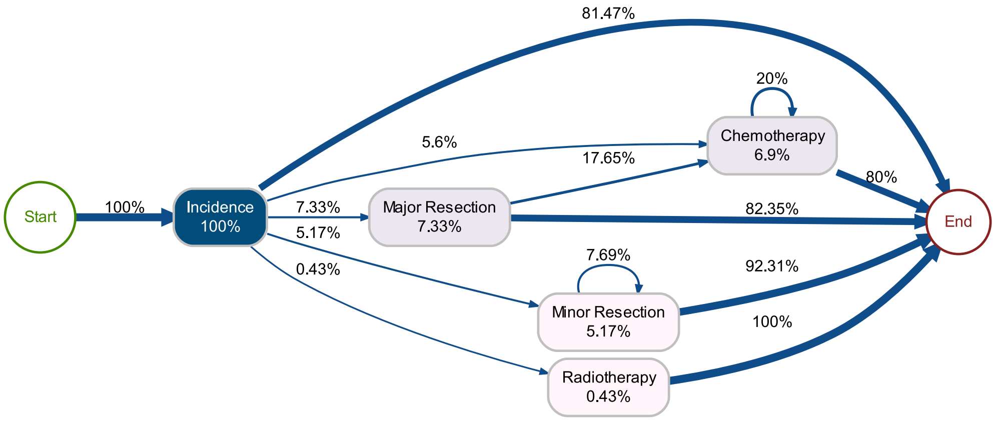

Figure 5. Treatment pathways for colon cancer patients with a Charlson comorbidity score of 0

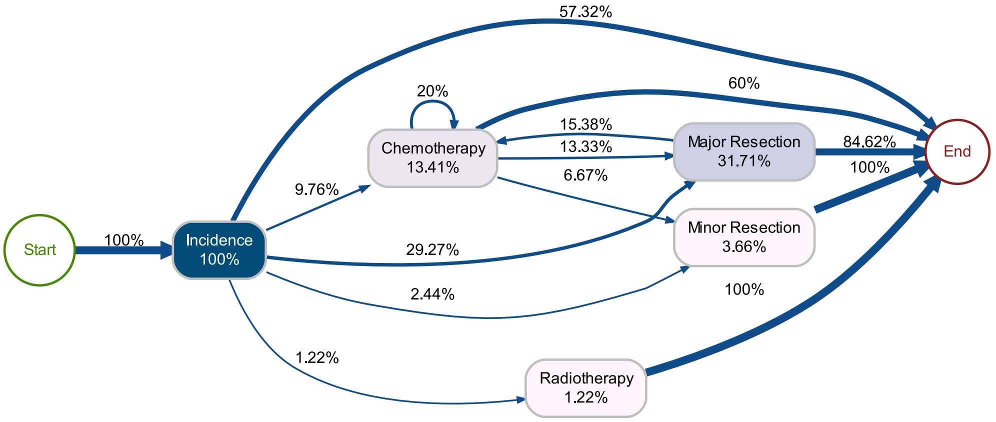

Figure 6. Treatment pathways for colon cancer patients in the most deprived quintile

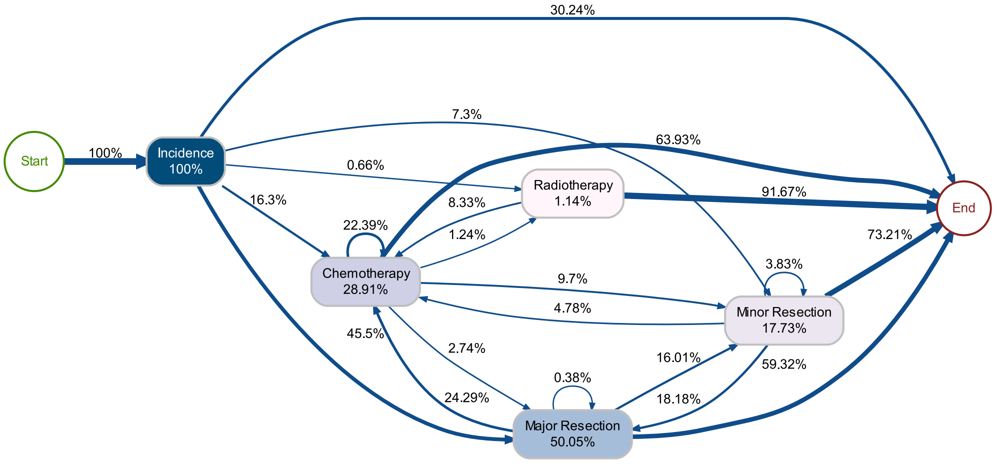

Figure 7. Treatment pathways for colon cancer patients in the least deprived quintile

Our method may be seen as an extension of the variant comparison approach set out in Cremerius et al. (2023), in which additional information in event data attributes are used to support the analysis of process variants. A care record will contain other, contextual information that may be used to divide the patient population into cohorts before events are selected, processes are generated, and variants are identified. Our analysis starts at a different point, with a different emphasis, but the underlying principle—that additional information can be incorporated within our analysis—is the same in each case.

An important limitation of our approach concerns the feasibility of obtaining high-quality contextual information, at scale, from care records. Each of the systems used within a hospital will have a different means of representing and accessing relevant data. While core demographic and treatment information may be retrieved from standard reports provided by these systems, data corresponding to indirect events—in particular, care plans—is difficult to extract and standardise. As Ingvar et al., 2021 observes, "while a number of proposed solutions for care plan support have been published and implemented, there is to date no consistency between them and often a very sparse description of the informatics concepts". We would recommend that attention is paid to the ongoing standardisation of additional, contextual information in care records: in particular, to the work underway in openEHR (Iglesias et al., 2022).

Another limitation is the lack of concrete guidelines for the presentation of findings and the interpretation at the end of each proposed iteration. The statistical analysis presented in Cremerius et al. (2023) may be helpful in this regard: the identification of measures that show a statistically-significant difference across variants could provide a basis for a more systematic approach and an indication of the potential value of further analysis based upon the same data. We will aim to address this limitation in future work.

The combination of process mining and epidemiology has the potential to accelerate healthcare transformation. The process mining method outlined in this paper is informed by epidemiology. There are clear opportunities for epidemiology, in turn, to be informed by process mining. At present, a typical epidemiological study involves checking to see whether receiving a particular treatment is associated with an improvement in outcomes (Morris et al., 2010). Using process mining, we can consider not only whether a particular treatment was received, but also when it was received, and whether it came before or after other treatments. We look forward to discussion and collaboration between epidemiologists and the process mining community.

Acknowledgements Authors supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1) (LC and OPD) and Elsevier (OPD).

References

Anwar, H., et al. (2011). Assessment of the under-reporting of diabetes in hospital admission data: A study from the Scottish Diabetes Research Network Epidemiology Group. *Diabetic Medicine*, *28*(12), 1514–1519. https://doi.org/10.1111/j.1464-5491.2011.03432.x

Aspland, E., et al. (2021). Modified Needleman–Wunsch algorithm for clinical pathway clustering. *Journal of Biomedical Informatics*, *115*, 103668. https://doi.org/10.1016/j.jbi.2020.103668

Baker, K., et al. (2017). Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. *International Journal of Medical Informatics*, *103*, 32–41. https://doi.org/10.1016/j.ijmedinf.2017.03.011

Batista, E., & Solanas, A. (2018). Process Mining in Healthcare: A Systematic Review. *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–6. https://doi.org/10.1109/IISA.2018.8633608

Cancer Research UK. (2015, May 15). *Bowel cancer treatment statistics*. Cancer Research UK. Retrieved June 8, 2023, from https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/diagnosis-and-treatment

Capita. (2014, September). *The quality of clinical coding in the NHS: Payment by Results data assurance framework*. https://www.chks.co.uk/userfiles/files/The_quality_of_clinical_coding_in_the_NHS.pdf

Charlson, M. E., et al. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, *40*(5), 373–383. https://doi.org/10.1016/0021-9681(87)90171-8

Chew, B.-H. (2019). Planning and Conducting Clinical Research: The Whole Process. *Cureus*. https://doi.org/10.7759/cureus.4112

CORECT-R. (2023). Retrieved August 24, 2023, from https://www.ndph.ox.ac.uk/corectr/corect-r

Cremerius, J., et al. (2022). Event log generation in MIMIC-IV research paper. *International Conference on Process Mining*, 302–314.

Cremerius, J., et al. (2023). Data-Based Process Variant Analysis. *Proceedings of the 56th Hawaii International Conference on System Sciences*. https://hdl.handle.net/10125/103031%20978-0-9981331-6-4

De Roock, E., & Martin, N. (2022). Process mining in healthcare – An updated perspective on the state of the art. *Journal of Biomedical Informatics*, *127*, 103995. https://doi.org/10.1016/j.jbi.2022.103995

dos Santos Silva, I. (1999). *Cancer epidemiology: Principles and methods*. International Agency for Research on Cancer.

Foley, T., & Vale, L. (2022). A framework for understanding, designing, developing and evaluating learning health systems. *Learning Health Systems*, *n/a*(n/a), e10315. https://doi.org/10.1002/lrh2.10315

Fox, F., et al. (2018). A Data Quality Framework for Process Mining of Electronic Health Record Data. *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 12–21. https://doi.org/10.1109/ICHI.2018.00009

Gasparini, A. (2018). Comorbidity: An R package for computing comorbidity scores. *Journal of Open Source Software*, *3*(23), 648. https://doi.org/10.21105/joss.00648

Gatta, R., et al. (2018). A Framework for Event Log Generation and Knowledge Representation for Process Mining in Healthcare. *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, 647–654. https://doi.org/10.1109/ICTAI.2018.00103

Goldacre, B., & Morley, J. (2022). *Better, Broader, Safer: Using health data for research and analysis. A review commissioned by the Secretary of State for Health and Social Care.* UK Department of Health and Social Care.

Iglesias, N., et al. (2022). Business Process Model and Notation and openEHR Task Planning for Clinical Pathway Standards in Infections: Critical Analysis. *Journal of Medical Internet Research*, *24*(9), e29927. https://doi.org/10.2196/29927

Ingvar, M., Blom, M. C., Winsnes, C., Robinson, G., Vanfleteren, L., & Huff, S. (2021). On the Annotation of Health Care Pathways to Allow the Application of Care-Plans That Generate Data for Multiple Purposes. *Frontiers in Digital Health*, *3*, 688218. https://doi.org/10.3389/fdgth.2021.688218

International Organization for Standardization. (2015). *Health informatics - System of concepts to support continuity of care (ISO 13940:2015)*. https://www.iso.org/standard/58102.html

Kaymak, U., et al. (2012). On process mining in health care. *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 1859–1864. https://doi.org/10.1109/ICSMC.2012.6378009

Kilpatrick, E. S., et al. (2022). Controversies Around the Measurement of Blood Ketones to Diagnose and Manage Diabetic Ketoacidosis. *Diabetes Care*, *45*(2), 267–272. https://doi.org/10.2337/dc21-2279

Kurniati, A. P., et al. (2018). Process mining in oncology using the MIMIC-III dataset. *Journal of Physics: Conference Series*, *971*, 012008. https://doi.org/10.1088/1742-6596/971/1/012008

Lear, R., et al. (2022). Patients' Willingness and Ability to Identify and Respond to Errors in Their Personal Health Records: Mixed Methods Analysis of Cross-sectional Survey Data. *Journal of Medical Internet Research*, *24*(7), e37226. https://doi.org/10.2196/37226

Lenz, R., & Reichert, M. (2007). IT support for healthcare processes – premises, challenges, perspectives. *Data & Knowledge Engineering*, *61*(1), 39–58. https://doi.org/10.1016/j.datak.2006.04.007

Martin, N., et al. (2020). Recommendations for enhancing the usability and understandability of process mining in healthcare. *Artificial Intelligence in Medicine*, *109*, 101962. https://doi.org/10.1016/j.artmed.2020.101962

Morris, E. J. A., et al. (2010). Surgical management and outcomes of colorectal cancer liver metastases. *British Journal of Surgery*, *97*(7), 1110–1118. https://doi.org/10.1002/bjs.7032

Munoz-Gama, J., et al. (2022). Process mining for healthcare: Characteristics and challenges. *Journal of Biomedical Informatics*, *127*, 103994. https://doi.org/10.1016/j.jbi.2022.103994

NHS Digital. (2021, June 22). *Clinical Classifications*. NHS Digital. Retrieved June 6, 2022, from https://digital.nhs.uk/services/terminology-and-classifications/clinical-classifications

Nouraei, S. A. R., et al. (2016). Accuracy of clinician-clinical coder information handover following acute medical admissions: Implication for using administrative datasets in clinical outcomes management. *Journal of Public Health*, *38*(2), 352–362. https://doi.org/10.1093/pubmed/fdv041

Oughtibridge, N. (2019, October 8). *A system of concepts for the continuity of care*. contsys.org. Retrieved August 27, 2023, from https://contsys.org/

Pijnenborg, P., et al. (2021). Towards Evidence-Based Analysis of Palliative Treatments for Stomach and Esophageal Cancer Patients: A Process Mining Approach. *2021 3rd International Conference on Process Mining (ICPM)*, 136–143. https://doi.org/10.1109/ICPM53251.2021.9576880

Remy, S., et al. (2020). Event log generation in a health system: A case study. *Business Process Management: 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings 18*, 505–522. https://doi.org/10.1007/978-3-030-58666-9_29

Rojas, E., et al. (2016). Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, *61*, 224–236. https://doi.org/10.1016/j.jbi.2016.04.007

Shrestha, A., et al. (2019). Quality of life versus length of life considerations in cancer patients: A systematic literature review. *Psycho-Oncology*, *28*(7), 1367–1380. https://doi.org/10.1002/pon.5054

Syed, R., et al. (2023). Digital Health Data Quality Issues: Systematic Review. *Journal of Medical Internet Research*, *25*, e42615. https://doi.org/10.2196/42615

Syriopoulou, E., et al. (2019). Understanding the impact of socioeconomic differences in colorectal cancer survival: Potential gain in life-years. *British Journal of Cancer*, *120*(11), 1052–1058. https://doi.org/10.1038/s41416-019-0455-0

Taylor, J. C., et al. (2021). Influence of age on surgical treatment and postoperative outcomes of patients with colorectal cancer in Denmark and Yorkshire, England. *Colorectal Disease*, *23*(12), 3152–3161. https://doi.org/10.1111/codi.15910