Investigating an Ontology-Informed Approach to Event Log Generation in Healthcare

Owen P. Dwyer[1][0000−0002−7157−6395], Lara Chammas[1][0000−0002−5605−2323], Emanuel Sallinger[1,2], and Jim Davies[1]

[1] University of Oxford
[2] TU Wien

Abstract. Process mining methods have the potential to reveal useful patterns in hospital data and provide insights into patterns of treatment, but healthcare has proven an exceptionally challenging domain due to the significant heterogeneity of possible events and the diversity of individual patients. When studying a patient's pathway, their history might contain hundreds of events unrelated to the disease of interest; determining which to include or exclude in an event log is difficult, and highly dependent on the situation. Typically, researchers only include the most common or correlated events (risking missing rare but significant exceptions), or hand-curate a list of events (slow, and difficult, especially where there might be thousands of possible events). We discuss an alternative approach that uses logical reasoning over ontologies to deduce plausibly related events, which might have implications in terms of time savings, the facilitation of reproducible research, and the enabling of comparisons across datasets and healthcare systems. We investigate the practical feasibility of such an approach by demonstrating it on real data, finding that the discovered processes approximate those produced through established approaches, with some caveats. The approach is a promising way to support conversations with domain experts and expedite human processes, and highlights the potential of structured domain knowledge to inform and enhance the event log generation process.

Keywords: preprocessing · filtering · ontologies.

1 Introduction

Clinical pathways describe the sequences of events that patients are expected to experience over the course of their journey through the healthcare system. The relationship between these ideal pathways and the real journeys experienced by patients, as observed in data from electronic health records (EHR), is an area of significant research interest, and process mining (PM) is a popular set of methods often proposed as a solution [19]. However, PM can struggle to produce useful insights from healthcare processes because they present several challenges: 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 communication and coordination between multiple specialists; and they are ad hoc, modified and interpreted according to individual preferences and professional judgement [20]. These challenges create data that is characterised by incomplete and noisy signals, high levels of process variation, and multitudes of exceptional behaviours that should ideally be captured rather than disregarded.

Given the complex state of raw healthcare data, preparing it for analysis is an intensive and time-consuming step, where modelling decisions and assumptions can have a significant impact on results; this is most evident when determining which clinical events to include in an event log. Human curation of events ensures a high degree of reliability, especially when domain experts are involved, but it is time-consuming and becomes increasingly challenging in a healthcare scenario where the possible events can easily number in the thousands. Statistical approaches, such as only considering the most common events, can be an easy way to simplify data for analysis. However, they risk overlooking rare but significant exceptions, or conversely including very common events that are wholly unrelated to the disease of interest.

In this paper, we describe a third approach to preprocessing for healthcare process mining: an ontological approach. Given that healthcare data is typically encoded in standardised terminologies, and that these terminologies are mappable to ontologies rich with semantics, we propose to use these ontologies to facilitate event log generation. By specifying domain-informed constraints, plausibly related events can be deduced, helping to reduce a large log of events to a smaller subset of more informative and pertinent ones.

Exploiting structured domain knowledge from ontologies has the potential to bring a number of improvements to the process of event log generation, such as:

- expediting the slow and labour-intensive process of codelist curation
- allowing codelists to be expressed as a set of intuitive, explainable rules
- providing rules that can be translated into dataset-specific terminologies, enabling reproducibility across healthcare systems and datasets
- promoting reproducibility and code sharing within the research community
- providing an approach available irrespective of a dataset's sample size

This paper therefore aims to investigate the extent to which event log generation can be automated through the use of ontologies. We first review previous work on preprocessing healthcare data, and the use of semantic reasoning in PM (Section 2). We outline an approach to event log generation (Section 3), and demonstrate it on real-world healthcare datasets, presenting our initial results (Section 4). Finally, we discuss the implications of these results and consider whether it effectively resolves the challenges identified, as well as explore the potential for future research (Section 5).

The approach presented here is inspired by, and aims to address, a number of the current challenges in healthcare PM, as established by Munoz-Gama et al. [19]. In particular, we address the ideas that there naturally exists high variability between patients; that infrequent behaviour should be valued; and that transparent approaches should be prioritised. Additionally, we also target

two of the current key challenges: *designing dedicated methodologies* by describing a healthcare-specific PM framework that leverages domain knowledge and facilitates comparison across healthcare providers; and *dealing with reality*, by demonstrating our approach on two different real-world datasets.

2 Background

In this section, we consider established approaches to preprocessing healthcare data for process mining, and introduce some of the sources of structured knowledge available in the healthcare domain.

2.1 Preprocessing in healthcare process mining

A 2021 literature review of preprocessing in PM [17] highlights event- and trace-level filtering as one of the most common preprocessing techniques used in the PM literature. However, this framework defines a filtering technique as one that determines the likelihood of the occurrence of events based on its surrounding behaviour, meaning it effectively focuses on the removal of logging mistakes. Determining whether a recorded event actually happened is a question of interest in healthcare PM, but given the very large size of a patient's record and the number of possible events, a more important question to answer first is whether a particular event — whether it happened or not — is even relevant to the particular disease or pathway being studied.

The obvious answer is to consider only the most common events in the study population. However, Tax et al. [22] demonstrate that filtering activities based on frequency alone does not fix problems of chaotic activities, and affects the quality of the final process model. It has been observed that "researchers and practitioners must go beyond simply filtering out infrequent behaviour from the event log" [19].

An alternative to this approach is human curation. In healthcare research, researchers studying health records typically create a *codelist*, a list of concepts in the dataset's terminology that are to be included for study, in order to rigorously define which exposures and outcomes are being investigated. The development of such a list requires three steps: a clear definition of the clinical feature of interest, a shortlist of the potential codes to include, and an iterative process of expert review [23]. This process is therefore naturally very time-consuming, and must be repeated for different datasets using different terminologies. Many researchers that hand-curate codelists in this way have created shared repositories of codes, in an attempt to establish standard and reproducible phenotypes for particular diseases. However, these libraries are focused around the phenotyping of diagnoses rather than procedures, with codelists largely consisting of ICD diagnostic codes [1,4]. Given that PM involves a similar process of deciding which procedural codes to extract from a healthcare record, the field of PM might benefit from an equivalent repository of procedure and event codelists, to similarly establish common phenotypes of disease pathways and facilitate reproducible research.

More recently, a standard approach to generating event logs from healthcare datasets has been proposed [9]: here, activity filtering involves a human identifying a codelist of procedures from those described in relevant medical guidelines. However, this approach is limited by the fact that the language of medical guidelines is typically vague. For example, the UK's guidance on colorectal cancer simply recommends that "surgery" be offered to patients: this does not represent a clear rule for which specific events should be included in a study, and will be interpreted differently by different healthcare systems and individuals [6]. Guidelines will also differ by location, making comparing results difficult. There is therefore a need to explore this activity selection stage in deeper detail, and establish clear and consistent rules for what is and is not included.

2.2 Ontologies, terminologies, and clinical coding

Healthcare procedures are represented in the electronic health record by clinical codes, which can follow one of many possible coding standards [11]. These standards allow for the conversion of medical records into terms that can then be statistically analysed, and provide a framework for standardisation between hospitals and countries. In this work, we encounter two major coding standards.

ICD is an international disease classification standard maintained by the World Health Organization. Its usage world-wide allows for comparability on rates of morbidity, including causes of death. ICD version 9 codes represent both diagnoses and procedures; ICD version 10 does not directly support procedure coding, but the USA maintains a local extension, the ICD-10 Procedure Coding System (ICD-10 PCS), which does.

OPCS is the procedure coding system used by hospitals within the UK's National Health Service. 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 sub-types. Strictly speaking, OPCS and ICD are *statistical classifications*, a subset of terminologies in which concepts are mutually exclusive and monohierarchical (every concept having exactly one parent), which aid in statistical reporting [11].

In comparison, **SNOMED-CT** — an international standard containing over 350,000 medical terms — describes semantic relationships and polyhierarchies amongst over 350,000 entities, making it closer to a true ontology. These can be reasoned over using logical languages such as the SNOMED Expression Constraint Language (ECL) [7]. This means that SNOMED-based logical reasoning can be a valuable source of knowledge, but real datasets still encode data in terminologies, meaning that terms need to be translated into their SNOMED equivalents for the benefits of a semantically rich ontology to be realised.

The combination of semantics and ontological reasoning with PM has been explored in the past. Medeiros and van der Aalst [18] describe the core idea of *semantic process mining* as the explicit relation or annotation of elements in a log with the concepts they represent, thereby making it possible to automatically reason or infer other relationships between them. They suggest that such ontologies might be useful in the log cleaning process. Other approaches

have used ontological structure to abstract events into less granular categories for analysis, or even to aggregate sequences of events into single events based on rules [16,21]. The difference between fine-grained events in EHR data and more abstract events in treatment guidelines has also been highlighted, with ontology-based abstraction proposed as a way to bridge this gap [15]. The use of ontologies in online systems has also been proposed, as a way of continuously verifying the plausibility of new events [12].

In summary, the preprocessing stages are a key part of the PM framework. Modelling assumptions and decisions, such as which events in a raw health record are included and which are discarded, will inevitably have an impact on the outcome of any analysis. The exact codelists and preprocessing steps used in studies are not always elaborated upon in detail in publications, which limits their reproducibility and usefulness. Cremerius et al. [9] have made some progress in rectifying this with a standard framework for preprocessing the popular MIMIC dataset. The remainder of this paper dives deeper into the event selection stages of this framework, considering how researchers might arrive at a rationale for the inclusion of individual events, and how ontologies might assist in this process.

3 Methods

Given that healthcare data is typically encoded in standardised terminologies, and that these terminologies are mappable to ontologies rich with semantics, we propose leveraging these properties and using domain knowledge to facilitate elements of the event log generation process. In our approach, a candidate list of procedure codes for analysis can be quickly generated by the execution of semantic rules over a domain ontology.

This paper aims to investigate the extent to which event log generation can be automated through the use of ontologies. To answer this question, we define several different possible sets of constraints that might be used to generate a list of candidate SNOMED concepts, we evaluate the resulting codelists in terms of their agreement with an established expert-curated codelist, using colon cancer as a case study, and we analyse the resulting processes in a real EHR dataset.

3.1 Logical constraints

Through an iterative process, we consider four candidate ECL queries for identifying concepts of interest in SNOMED, listed in table 1. Firstly, to retrieve all concepts that might be plausibly related to colon cancer, we formalise the question "which concepts are chemotherapy, radiotherapy, or are procedures that occur on the colon or part of the colon?" into a set of rules (Query 1).

This initial query retrieved many codes which were only tangentially colon-related: for example, G02.4, *total oesophagectomy and interposition of microvascularly attached colon* is unlikely to be related to colorectal cancer as the oesophagus is the primary site, and the colon is of secondary concern. Therefore,

1	(< 71388002|Procedure| : 405813007|Procedure site - Direct| = << (* : R 363698007|Finding site| = <<363406005)) OR << 367336001 | Chemotherapy (procedure) | OR << 108290001 | Radiation oncology AND/OR radiotherapy (procedure) |
2	(< 71388002|Procedure| : (405813007|Procedure site - Direct| = << (* : R 363698007|Finding site| = <<363406005)) AND ([1..1] 405813007|Procedure site - Direct| = *)) OR << 367336001 | Chemotherapy (procedure) | OR << 108290001 | Radiation oncology AND/OR radiotherapy (procedure) |
3	(< 71388002|Procedure| : 405813007|Procedure site - Direct| = <<71854001) OR << 367336001 | Chemotherapy (procedure) | OR << 108290001 | Radiation oncology AND/OR radiotherapy (procedure) |
4	(< 71388002|Procedure| : (405813007|Procedure site - Direct| = <<71854001) AND ([1..1] 405813007|Procedure site - Direct| = *)) OR << 367336001 | Chemotherapy (procedure) | OR << 108290001 | Radiation oncology AND/OR radiotherapy (procedure) |

Table 1. SNOMED ECL queries 1-4

we introduced an additional constraint that procedures must have one and only one site (Query 2).

Query 1 also retrieved some procedures not related to the colon at all, for example H34.3, *open cryotherapy to lesion of rectum*. These arise because the SNOMED concept representing *malignant neoplasm of colon* has child concepts including *overlapping malignant neoplasm of colon and rectum*, so the rectum is introduced as a procedure site of interest. Therefore, we created a more specific query: instead of any possible site for any colon cancer concept, we consider only the procedures on the colon, or a descendant thereof (Query 3).

Finally, we combine the additions from Queries 2 and 3 to constrain our output to those procedures with exactly one procedure site and a procedure site that is a descendant of the colon, or the colon itself (Query 4).

3.2 Analysis

To benchmark the usability of logic reasoning for clinical activity selection, and to evaluate the retrieval performance of Queries 1-4, we compare their result sets to a reference list of colorectal cancer treatments, as curated by domain experts. Specifically, we use a codelist from the COloRECTal cancer Repository (CORECT-R) project [10]. The project's codelist [5] maps 197 different OPCS codes to five different treatment categories: major resection, minor resection, stoma, stent, and bypass. Furthermore, to investigate the entire patient journey for colorectal cancer, we extended this code list, in consultation with subject matter experts, to include additional codes for diagnostic tests, chemotherapy, radiotherapy, and surgical procedures related to the maintenance of treatments, amounting to a total of 497 OPCS codes*. Each ECL query is evaluated 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). Following this, we examine the resulting process model from our derived codelist, and compare it to one generated from our benchmark list.

* available at `https://github.com/larachammas/EventLists`

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

Table 2. Query precision and recall against original and extended CORECT codelists

3.3 Data

We demonstrate this methodology by applying it to two separate real-world datasets. Firstly, we use data collected by Oxford University Hospitals (OUH) NHS Foundation Trust, which operates four hospitals in Oxfordshire, south-east England. In this dataset, procedures are coded using OPCS. We also make use of the MIMIC-IV dataset [13,14], a freely accessible EHR dataset from Massachusetts, USA which has been widely used in previous PM studies. This dataset uses a mixture of ICD-9 and ICD-10-PCS procedure codes.

We use the official NHS maps for converting SNOMED codes to their corresponding OPCS codes [8], which are intended to assist clinical coders in converting codes for billing purposes. Most SNOMED concepts actually map to multiple OPCS codes — either a choice of multiple codes, or a mandatory set of multiple codes. Given that the aim of this study is to establish plausibility rather than to guarantee causality, we assume that any OPCS code that is mapped in any way to a SNOMED concept could plausibly be related to it.

4 Results

In this section, we present the initial results of our proposed approach. Firstly, we compare the quality of the generated codelists against our reference list. We then compare the process models generated using these lists, and discuss differences and similarities. Finally, we examine the distribution of codes in an additional dataset to see how a statistical element might add value to our approach.

4.1 Codelist quality

Table 2 compares each of the four queries in terms of their precision and their recall, against both the original and extended CORECT-R lists. OPCS codes in chapters Y and Z — supplementary codes that describe additional information such as sites or approaches — are excluded since as a rule they never appear in the primary position in records, and therefore have little bearing on PM results.

Query 1 exhibits high recall, retrieving 85% of the original CORECT-R codes, but a low precision (26%), meaning that many extra codes were retrieved. Conversely, Query 4, the most restrictive, has the highest proportion of CORECT-R codes amongst its results (51%), but captures a much smaller set of the whole list (51%). In the extended codelist, precision is naturally higher and recall lower

than in the original list, since more of the retrieved codes are in the reference list, but a smaller proportion of the entire list is retrieved. A consistent pattern is that increasing the constraints on the original ECL query can significantly increase precision, but at a cost of much lower recall.

Figure 1 shows the recall of each query, further broken down by the concepts' categories. Concepts representing blood tests, diagnostic testing and patient assessment were never retrieved, presumably because they are not disease specific and therefore do not have a clear site; similarly, endoscopy of the upper gastrointestinal tract was not retrieved as it does not have a relevant procedure site, but it is commonly included in colon cancer analyses as a common route to diagnosis. Recall was highest in surgical topics, since they were well annotated with procedure sites in the ontology. The differences in constraints between queries 1-4 only had a noticeable affect on the *resection* and *other operations* codes.

4.2 Processes

To investigate the usefulness of these codelists in PM, we generate two directly follows graphs (DFGs), one using the extended CORECT-R codelist and one using the Query 1 codelist, chosen as it was the list with the highest overlap with CORECT-R. In these graphs (Figure 2), 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. The codes in our new codelist were grouped onto a smaller number of categories for visualisation based on the OPCS codes' text descriptions (e.g. 'resection' encompasses all codes mentioning 'resection', 'excision', 'lesion' or 'exentoration'). For clarity and simplicity, *minor* and *major* resection were grouped into one *resection* category, as were *lower GI* and *upper GI endoscope*. 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 (58.51% in our codelist, 53.89% according to the CORECT codelist), indicating that the derived codelist arrived on a broader definition of what constitutes a resection. Both processes share similarities: the most common first event was resection; imaging and radiotherapy were typically

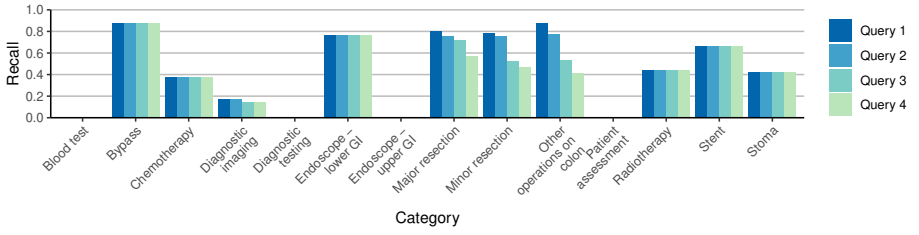

Fig. 1. Recall of queries by category compared to the extended CORECT reference set

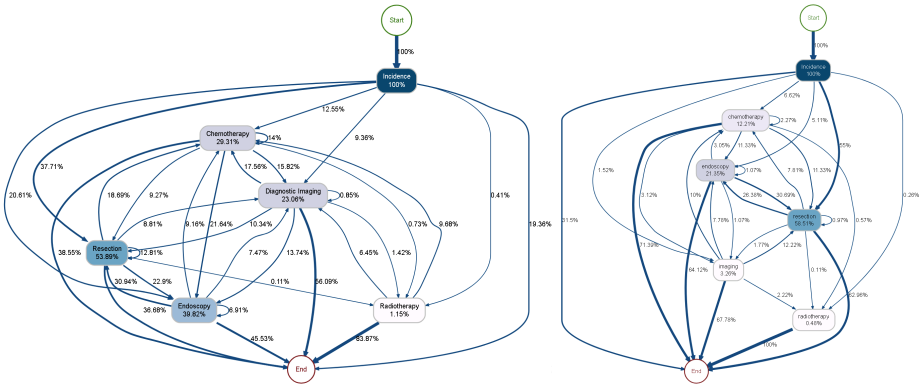

Fig. 2. Directly-follows graphs for Query 1 (left) and CORECT-R (right) codelists

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. These differences will in turn affect the entire process model. In particular, this potentially explains why chemotherapy, endoscopy, resection and imaging were so often the last event in a pathway: in actuality, 21% of chemotherapies are followed by endoscopies, 15% by imaging, and 20% by resection, but an extra 17% of patients had chemotherapy, and an extra 20% had imaging, which were never properly identified in the first place. Therefore, whilst our generated codelist does in some aspects provide a rough sketch of patient journeys, it is certainly far from complete.

4.3 Statistics and the MIMIC dataset

A limitation of the OUH dataset used here is that the data available at this point only consists of patients with colorectal cancer. An ontological approach could potentially be enhanced by combining both logical and statistical knowledge, enabling the identification of codes that do not have a clear logical link, but are nonetheless correlated with colon cancer. To test this, we examined the MIMIC dataset, which contains data on an entire cohort of patients regardless of disease history.

We divided the MIMIC patients into a control cohort (those without a colon cancer (ICD-10 C18) diagnosis, $n = 120,906$) and a case cohort (those with the diagnosis, $n = 1,046$), and calculated how frequently each code appeared in each group. The codes with the greatest difference in frequency between the two patient groups that did not have a known counterpart in Query 1's codelist [3,2] are listed in Table 3. This immediately raises some interesting results: *large-to-large intestinal anastomosis* and *other small-to-large intestinal anastomosis* are highly correlated with colorectal cancer diagnosis, and are obviously colon related, but are not in our codelist. More suprisingly, *laparoscopic right hemi-*

ICD version	ICD concept	control frequency (%)	case frequency (%)	odds ratio	P
9	**17.35** Laparoscopic left hemicolectomy	0.04	2.49	70.02	$1.06e^{-35}$
9	**45.93** Other small-to-large intestinal anastomosis	0.15	7.65	55.54	$1.39e^{-99}$
9	**45.94** Large-to-large intestinal anastomosis	0.08	2.87	34.97	$1.27e^{-33}$
10	**0D1B0Z4** Bypass Ileum to Cutaneous, Open Approach	0.18	2.58	15.08	$6.71e^{-22}$
9	**45.25** Closed [endoscopic] biopsy of large intestine	1.14	11.28	11.05	$2.71e^{-74}$

Table 3. The top five MIMIC-IV procedure codes with the greatest differences in frequency between colon cancer and non-colon cancer patients. Considered were codes that appeared > 20 times in the case cohort, were more common in case than control patients, and statistically significant ($P < 0.05$) according to Fisher's exact test.

colectomy does appear in our codelist, but *laparoscopic left hemicolectomy* does not. An examination of the mapping files reveals that left hemicolectomy has no mapping to any SNOMED concept, so regardless of whether its SNOMED counterpart was identified in our search, it would not be detected in an ICD-9 encoded dataset.

5 Discussion and Conclusions

These results establish that a logical query over SNOMED's domain knowledge is capable of approximating a human-prepared codelist with some accuracy, but there exists a clear trade-off between precision and recall: either a relatively complete codelist is retrieved with a large number of extra codes, or an incomplete codelist is retrieved in which every code has as high chance of being known to be relevant. The differences between process models produced via this approach and from a traditional curated codelist support this. Additionally, challenges appeared in the form of grouping codes into categories. The original approach considered was to group codes according to relationships in the SNOMED ontology - for example, all codes that were a descendent of "imaging" could safely be considered imaging. However, in a high proportion of cases, many possible SNOMED concepts mapped onto one OPCS code creating multiple possible categories. An additional problem is the need for these categories to be aligned with the ones from the benchmark codelist to allow for a useful comparison. Hence, the solution of grouping codes by text was arrived upon.

The supplementary analysis of the the MIMIC dataset indicated that many codes existed that were clearly correlated with colon cancer, but not identified by our proposed approach. In addition, some unusual results highlighted gaps in the mappings' coverage. Overall, these results suggest that the approach proposed here is, whilst an interesting experiment, far from a replacement for discussion and input from domain experts, and in fact underline their importance.

There are some potential advantages to the presented approach. It does provide a reasonably accurate list on which to begin discussions, speeding up the initial stages of the development process. It allows the majority of concepts in

a codelist to be described simply, as a set of intuitive, explainable rules. It in principle allows the description of codelists in terms of a well-known ontology that can be translated into the specific languages used by individual datasets, although this is highly dependent on both the completeness of the original ontology and the quality of mappings available. It also provides a viable approach for applications where datasets are very small, rendering statistical examinations of the most common codes unhelpful. Finally, the specification of desired process events in terms of their relationship to a particular target disease keeps PM focused on a concrete health issue, ensuring clinical meaningfulness and alignment between PM studies and research questions around healthcare delivery.

It also highlights several avenues for future research; in particular, further work is required to establish the extent to which this approach is generalisable to other diseases, and whether it might also work in other application areas outside of healthcare with their own ontologies.

In summary, we have found that domain knowledge, even in the form of established ontologies, is far from perfect and cannot instantly automate the process of event log preparation. However, it can provide a meaningful framework to generate initial ideas to be iterated upon. This study highlights that no single approach — whether reliant on ontologies, conversation with domain experts, or pure statistics — can fully capture the complex nature of healthcare processes, and that a combination of all are required to fully explore health processes and answer questions around modern healthcare delivery.

Acknowledgements

OPD and LC are supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1). OPD is supported by Elsevier. ES is supported by the Vienna Science and Technology Fund (WWTF) (10.47379/ICT2201, 10.47379/VRG18013, 10.47379/NXT22018); and the Christian Doppler Research Association (CDG) JRC LIVE.

References

1. HDRUK Phenotype Library, https://phenotypes.healthdatagateway.org/
2. ICD-9-CM Diagnostic Codes to SNOMED CT Map, https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html
3. ICD-9-CM to and from ICD-10-CM and ICD-10-PCS General Equivalence Mappings, https://www.nber.org/research/data/icd-9-cm-and-icd-10-cm-and-icd-10-pcs-crosswalk-or-general-equivalence-mappings
4. OpenCodelists, https://www.opencodelists.org/
5. CORECT-R Data Coding v1.0 (Oct 2020), https://www.ndph.ox.ac.uk/corectr/files/corect-r-data-coding-v1-0-oct20.pdf
6. Colorectal cancer (NICE guideline NG151) (Dec 2021), https://www.nice.org.uk/guidance/ng151/resources/colorectal-cancer-pdf-66141835244485
7. Expression Constraint Language - Specification and Guide Version 2.1 (Aug 2022), http://snomed.org/ecl

8. SNOMED CT to ICD-10 and OPCS-4 Map Table Technical Specification and Implementation Guidance (UK Edition) (Feb 2023), https://nhsengland.kahootz.com/gf2.ti/f/762498/159916325.1/PDF/-/doc_SnomedCTICD10OPCS4MapTableTechnicalSpecificationImplementationGuidance_UK%20Edition__Current-e.pdf
9. Cremerius, J., Pufahl, L., Klessascheck, F., Weske, M.: Event Log Generation in MIMIC-IV. In: Process Mining Workshops (ICPM 2022). LNBIP, vol. 468, pp. 302–314. Springer (2023)
10. Downing, A., Hall, P., Birch, R., Lemmon, E., Affleck, P., Rossington, H., Boldison, E., Ewart, P., Morris, E.J.A.: Data Resource Profile: The COloRECTal cancer data repository (CORECT-R). Int. J. Epidemiol. **50**(5), 1418–1418k (Oct 2021)
11. Haendel, M.A., Chute, C.G., Robinson, P.N.: Classification, Ontology, and Precision Medicine. N. Engl. J. Med. **379**(15), 1452–1462 (Oct 2018)
12. Helm, E., Buchgeher, G., Ehrlinger, L.: Online Plausibility Checks for Patient Pathways with Medical Ontologies. In: PODS4H at ICPM (Posters) (Oct 2022)
13. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L., Mark, R.: MIMIC-IV (version 1.0) (2020). https://doi.org/10.13026/s6n6-xd98
14. Johnson, A.E.W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., Lehman, L.w.H., Celi, L.A., Mark, R.G.: MIMIC-IV, a freely accessible electronic health record dataset. Sci. Data **10**(1), 1 (Jan 2023)
15. Klessascheck, F., Lichtenstein, T., Meier, M., Remy, S., Sachs, J.P., Pufahl, L., Miotto, R., Boettinger, E., Weske, M.: Domain-Specific Event Abstraction. Business Information Systems pp. 117–126 (Jul 2021)
16. Leonardi, G., Striani, M., Quaglini, S., Cavallini, A., Montani, S.: Towards Semantic Process Mining Through Knowledge-Based Trace Abstraction. In: Data-Driven Process Discovery and Analysis. LNBIP, vol. 340, pp. 45–64. Springer (2019)
17. Marin-Castro, H.M., Tello-Leal, E.: Event Log Preprocessing for Process Mining: A Review. Appl. Sci. **11**(22), 10556 (Jan 2021)
18. Alves de Medeiros, A.K., van der Aalst, W.M.P.: Process Mining towards Semantics. In: Advances in Web Semantics I: Ontologies, Web Services and Applied Semantic Web, pp. 35–80. LNCS, Springer (2009)
19. Munoz-Gama, J., Martin, N., Fernandez-Llatas, C., Johnson, O.A., Sepúlveda, M., Helm, E., Galvez-Yanjari, V., Rojas, E., Martinez-Millana, A., Aloini, D., Amantea, I.A., Andrews, R., Arias, M., Beerepoot, I., Benevento, E., Burattin, A., Capurro, D., Carmona, J., Comuzzi, M., ... Zerbato, F.: Process mining for healthcare: Characteristics and challenges. J. Biomed. Inform. **127**, 103994 (Mar 2022)
20. Rebuge, Á., Ferreira, D.R.: Business process analysis in healthcare environments: A methodology based on process mining. Inf. Syst. **37**(2), 99–116 (Apr 2012)
21. Remy, S., Pufahl, L., Sachs, J.P., Böttinger, E., Weske, M.: Event Log Generation in a Health System: A Case Study. In: Business Process Management. pp. 505–522. LNCS, Springer (2020)
22. Tax, N., Sidorova, N., van der Aalst, W.M.P.: Discovering more precise process models from event logs by filtering out chaotic activities. J. Intell. Inf. Syst. **52**(1), 107–139 (Feb 2019)
23. Watson, J., Nicholson, B.D., Hamilton, W., Price, S.: Identifying clinical features in primary care electronic health record studies: methods for codelist development. BMJ Open **7**(11), e019637 (Nov 2017)