Reasoning over Health Records with Vadalog: a Rule-based Approach to Patient Pathways

Owen P. Dwyer[1], Teodoro Baldazzi[2,4], Jim Davies[1], Emanuel Sallinger[3,1] and Adriano Vlad[4,1,3]

[1]*University of Oxford, Oxford, United Kingdom*
[2]*Università Roma Tre, Rome, Italy*
[3]*TU Wien, Vienna, Austria*
[4]*Prometheux, London, United Kingdom*

Abstract
In recent years, the scale of biomedical and healthcare data has grown exponentially, leading to companies building large enterprise knowledge graphs as well as scalable and intelligent processing systems to exploit them. In this high-stakes domain, the transparency of data-driven processes is paramount to ensure high levels of trustworthiness and accountability for patient safety. This requirement has acted as catalyst for a rising interest in deductive approaches that use expressive declarative languages to represent domain knowledge, as well as powerful logic-based reasoning systems for the highly efficient and explainable deduction of new information. In this work, we explore the topic of patient pathways. This perspective on health records is a key concept in modern healthcare, but is not naturally evident from raw data, requiring data modelling decisions and domain expertise to explore in depth. We explore the utility of declarative approaches in deriving pathways for groups of patients from health records, and consider how these rules can aid in the intuitive interpretation and explanation of healthcare data. We employ Vadalog, a highly expressive language for knowledge representation and reasoning, to formulate tasks as logical rules, and use our state-of-the-art reasoning framework to achieve full transparency and explainability throughout the inference process, demonstrating these principles on a publically available dataset. This research strives to bridge the gap between the biomedical domain and ontological reasoning methodologies, paving the way for the future use of declarative approaches to facilitate population studies, precision medicine, and more transparent and explainable approaches to health data science.

Keywords
Ontological reasoning, Vadalog, Healthcare, Patient pathways

1. Introduction

As the scale of healthcare data continues to grow, so do the opportunities to analyse it and discover valuable insights into the realities of patient care. There is therefore a great need for scalable and effective data processing systems that can quickly turn vast volumes of raw data into valuable knowledge. This paper is motivated by the very concrete challenges faced by health systems worldwide. In particular, it is driven by our involvement in national cancer data

projects in the UK, together with our partners within the UK's National Health Service and others. The challenge we raise — both as our concrete problem at hand, as well as a challenge for the declarative AI community — is that of clinical pathways, which we will introduce in detail in this work. Our specific research interests are in colorectal cancers, but the principles discussed here are generalisable and relevant for a wide range of diseases.

The current computational landscape is dominated by inductive machine learning approaches. While these are very effective for many applications, they often exhibit a lack of transparency, appearing as black boxes and providing very limited insight into the reasoning behind their predictions. This is a particularly challenging drawback in healthcare contexts, where transparency is a vital step in building public trust in such systems. A second apparent difficulty consists of expressing complex domain-specific problems, as conceived by healthcare domain experts, into a computable formalisation that enables the efficient data analysis from AI to actually be leveraged. Declarative languages offer a potential solution to both problems, allowing the expression of complex queries in clear logical terms, understandable and authorable directly by domain experts [1].

Healthcare data is far from homogenous, and different scenarios call for different approaches to its analysis and interpretation. In this work, we focus on a particularly relevant one, namely clinical pathways. Clinical pathways are structured plans based on the best available evidence, which aim to optimise patient outcomes, minimise errors, reduce costs, and streamline the delivery of care across different settings. They play a crucial rule in standardizing care, improving efficiency and resource utilization, promoting evidence-based practice, ensuring continuity of care, and facilitating monitoring and quality improvement. However, they also involve complex networks that combine vast amounts of technical terminology, and require real-world data to be effectively monitored and understood. They are not naturally evident from raw electronic health record (EHR) data, requiring a series of assumptions, data modelling decisions, and domain expertise to explore in depth.

This paper explores such challenges and proposes the use of rule-based approaches to address them. Traditionally, patient pathways have been approached within the healthcare domain, with limited engagement from the knowledge and rules community. However, the complexities involved in modeling and interpreting patient pathways demand a more comprehensive and flexible approach. Rule-based methodologies offer a promising solution by enabling flexible representation and reasoning. Specifically, we choose the Vadalog language [2] and framework for its high expressive power in knowledge representation and its strong ontological reasoning capabilities. This technology has been successfully applied to many use cases and industrial applications, in particular finance [3, 4, 5, 6, 7, 8], and biomedicine [1]; in this work we demonstrate how it can be effective in the specific context of patient pathways. By formulating rules in Vadalog and performing reasoning with the associated framework, pathway questions can be quickly and easily answered in a fully transparent and explainable way.

We demonstrate the power of Vadalog in this context through three relevant scenarios. First, we explore how rule-based approaches can effectively aggregate and filter raw EHR data into an interpretable form that aids the answering of pathway-related questions, and how the integration of raw data with domain ontologies expedites this process. Then, we describe how a relatively simple graph traversal process can map patients' journey and help to answer key healthcare research questions. Finally, we discuss how additional unstructured data from

free-text sources can be effectively combined with the observational data to generate new insights into patient care processes.

By leveraging the logical formalism offered by Vadalog, we address the inherent complexity of patient pathways. Our findings emphasize the relevance and importance of rule-based approaches, and Vadalog's capabilities in understanding and interpreting patient pathways within EHR data. This work contributes to the advancement of knowledge in the field and lays the groundwork for further research in this critical area.

Overview The rest of the paper is organised as follows. In Section 2 we discuss the importance and key challenges of clinical pathways. In Section 3, we present some relevant background on Vadalog reasoning. In Section 4 we illustrate how we employ Vadalog to power the use cases. Finally, in Section 5 we discuss our results and their implications for future work in this area.

2. Electronic health records and patient pathways

In healthcare, clinical pathways are guidelines that aim to standardise care for a particular condition or group of patients. They describe the sequences of events that patients should experience over the course of their journey through the healthcare system, providing a recommended "route" through identification, diagnosis, treatment and follow-up. Recommended pathways are widely used within the UK's National Health Service, and they exist at a variety of different levels, from the guidelines set out by national standards bodies to the multitude of local interpretations designed to meet the needs of individual providers.

Despite the status of these official pathways that aim to standardise care, there will always be variations in real clinical practice. Indeed, there is an expectation that clinicians should respect patient autonomy and preferences, and that they should offer treatments based on each individual case and their professional opinion and experience. Therefore, there will always be some degree of variation from any recommended pathway. Whilst clinical pathways generally reduce complications, length of stay, and costs [9], they are ultimately only as good as the evidence they are based on, and effectiveness can vary considerably within distinct patient subgroups [10]. Thus, there exists a variety of open research questions surrounding their use, many of which rely on effective patient stratification to be answered. As healthcare increasingly adopts the precision medicine perspective — tailoring treatments to individual patients — the ability to measure the effectiveness of individual pathways, and the extent to which it varies based on clinical factors and demographic characteristics, will become increasingly important.

Overall, the widespread use of clinical pathways, combined with the inevitability of variations from the recommended route, prompts a number of research questions. How many patients actually follow the recommended or expected pathway for their condition? What are the most common real-world pathways? Most importantly, is alignment with a recommended pathway actually associated with positive outcomes? The increasing availability of large databases of electronic health records (EHR) means that we can begin to answer these questions on a large scale, and the pathways taken by individual patients can be analysed, stratified, and compared to identify distinct subgroups in terms of treatment. This might involve either instances of sub-optimal care that contribute to poor outcomes, or local practices that actually lead to improved ones.

2.1. Reasoning over electronic health records

Graph-based data models, and in particular knowledge graphs (KGs), are an appealing way to represent biomedical knowledge for several reasons, having a number of properties which make them well-suited to represent and analyse complex healthcare data. 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. Their structures support both logic-based reasoning and machine learning approaches for analysis, and they support the merging of both observational data and domain knowledge which is often required to answer complex questions.

For these reasons, many authors have explored the application of graph-based reasoning to biomedical and healthcare contexts. For example, Alfonsi et al. present a data model that combines viral sequence data with references in scientific literature, demonstrating how complex questions that span the two can be simply expressed with Vadalog [1].

Whilst knowledge graph reasoning has been applied to a wide range of biomedical problems, as of yet relatively few attempts have been made to use KG formalisms to solve pathway-related problems. Several authors have discussed the general advantages of logical languages and graph formalisms for EHR data in general: both Stothers et al. [11] and Yoon et al. [12] find that queries over patient data are syntactically simpler and execute faster in Neo4j than in PostgreSQL and MySQL respectively. Campbell et al. [13] demonstrate an approach that combines domain knowledge from the SNOMED-CT terminology with observational data from patient records, and find that a graph implementation allows for complex queries that would traditionally require many steps in traditional DBMS. Similarly, Piro et al. [14] encode rules surrounding diabetes care, and are able to reduce 3000 lines of SQL code down to 174 logical rules. Therefore, it is clear that rule-based reasoning can add significant value to EHR data.

Some research has explored the encoding of clinical guidelines into computational form, known as *computer-interpretable guidelines* (CIGs). However, attention has generally focused on encoding the individual decision points in a healthcare process, for the benefit of clinicians using decision support systems. This is a different problem from our interpretation of patient pathways, which is interested in tracing long-term processes [15]. We refer to a definition of clinical pathways that aligns with the ISO standard for continuity of care: a general plan, often set out at a national level, that reflects best practice and is broadly applicable to all patients [16].

Overall, the idea of the patient pathway is widely used in modern healthcare, and is effectively a set of rules that describes expected or recommended sequences of events. Bridging the gap between these hypothetical scenarios and the data that describes how they function in reality is a challenge of great interest and importance. This data is complex, multimodal, and often relies on references to healthcare ontologies to be properly interpreted: properties which make it a natural fit for logic-based reasoning.

3. Reasoning with declarative languages

To address these challenges, we propose a deductive approach powered by state-of-the-art logic-based techniques to reason over large KGs and achieve scalable and explainable analysis.

To guide our discussion, we first introduce some preliminary concepts.

Knowledge graphs At the core of our solution is a *Knowledge Graph Management System* (KGMS), a middleware that enables ontological reasoning on knowledge graphs (KGs). These are a semi-structured data model for knowledge representation and reasoning composed of (i) an extensional component, i.e. existing entities and relations integrating knowledge from heterogeneous data sources; (ii) an intensional component, i.e. a set of logic rules describing domain knowledge; and (iii) a derived extensional component produced via the application of the logic rules to the extensional component.

The Vadalog language Our framework employs Vadalog [2], a declarative language for ontological reasoning which enables us to effectively model a wide range of real-world problems with a readable and concise syntax, without the need to program complex control flows or to design algorithms. Vadalog expresses problems at a high level, empowering domain experts to act as data analysts [17]. Vadalog is based on Warded Datalog$^\pm$, a fragment of the Datalog$^\pm$ family of languages [18]. It encompasses plain Datalog (thus incorporating full recursion, essential for graph navigation tasks) and allows existential quantification (for instance, to address clustering settings). At the price of very mild syntactic restrictions, it guarantees PTIME data complexity for query answering. Furthermore, Vadalog extends Warded Datalog$^\pm$ with relevant features of practical utility, such as monotonic aggregation [19], selection conditions, and algebraic operations. With these advanced tools, Vadalog is capable of efficiently encoding graph traversal algorithms; it captures both regular path queries for navigating graphs using pattern matching (e.g. Cypher), and SPARQL under the OWL2 QL regime, for querying the semantic web.

A Vadalog program consists of a set of facts and *tuple-generating dependencies* (TGDs), i.e., first-order sentences of the form $\forall \bar{x} \forall \bar{y}(\varphi(\bar{x}, \bar{y}) \rightarrow \exists \bar{z}\ \psi(\bar{x}, \bar{z}))$, where the *body* $\varphi(\bar{x}, \bar{y})$ and the *head* $\psi(\bar{x}, \bar{z})$ are conjunctions of atoms over the respective predicates, $\bar{x}, \bar{y}$ are vectors of universally quantified variables and constants, and $\bar{z}$ is a vector of existentially quantified variables. Quantifiers can be omitted, and conjunction is denoted by a comma.

4. Applying Vadalog to patient pathway problems

Having introduced the application area of clinical pathways and its importance in the modern healthcare landscape, as well as the history of logical reasoning applied to healthcare data, we now outline three use cases in which we answer pathway questions with a logical reasoning approach. The combination of components presented here forms a knowledge graph: we apply the presented Vadalog rules — the intensional component — to reason over the MIMIC dataset — the extensional component — and extract new knowledge — the derived extensional component. In our first use case, we outline how a typical pre-processing pipeline for preparing a patient's event log can be implemented as Vadalog rules. We also describe how the process of aggregating and filtering the observational data from health records can be automated and enhanced when combined with rules from domain ontologies. The second use case considers the graph exploration process through recursion and aggregation functions in Vadalog, and relates this to current questions in the population health literature. Finally, we introduce additional contextual data generated from the scientific literature, and show how this information can
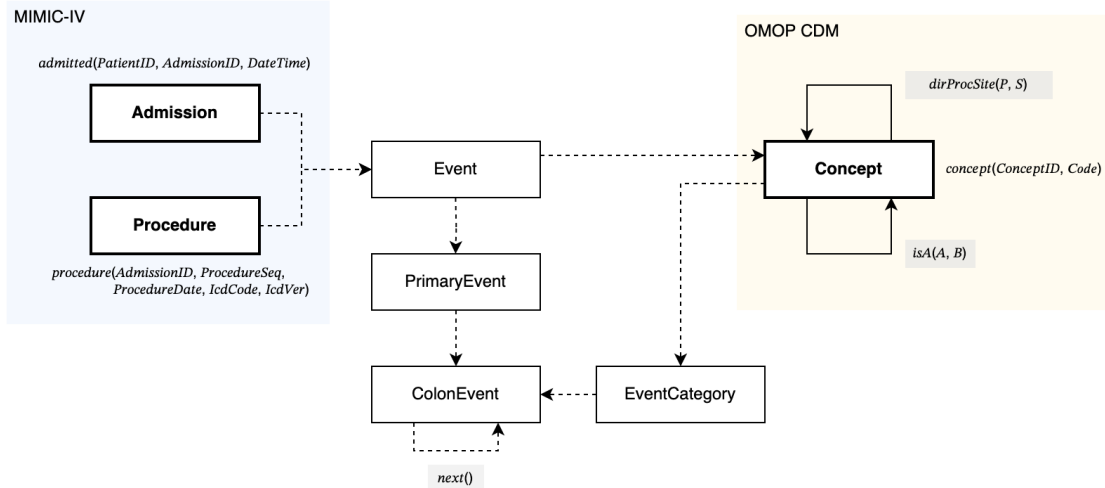
Figure 1: The data model described in this case study, including both the ground-truth facts included in the initial datasets (bold), and the additional structure derived

be combined with our data to generate new insights. The data used for the use cases is public, whereas the Vadalog reasoning framework will be made available upon request.

4.1. Building timelines

For demonstration purposes, we use the MIMIC-IV dataset, a freely accessible set of electronic health records from the Beth Israel Deaconess Medical Center in Massachusetts, USA [20, 21]. In the subset of data used for this demonstration, we consider the *admissions* and *procedures* tables: our data follows a roughly hierarchical structure, with each patient taking part in one or more admissions, and each admission containing one or more procedures. Each procedure is characterised by a code from the International Classification of Diseases (ICD) terminology, but the dataset uses a mixture of ICD versions 9 and 10, meaning that both the code and the version number are required to fully interpret an event. Since patient pathways look vastly different for every condition, this demonstration focuses on one disease area: we only consider the 1,046 patients having a recorded diagnosis of colon cancer – i.e. patients for whom the ICD-10 code C18 or the ICD-9 code 153 is present in the *diagnoses* table.

As well as an ICD code that characterises the clinical procedure, each procedure is also manually assigned a sequence number by a clinical coder, representing that procedure's relative importance in the admission [22]. It is common practice in data-driven studies to summarise an admission by considering only the primary procedure code as representative of the admission [23]. Thererore, to summarise the data into a form ready for easy analysis, we define an *event* as a procedure undergone by the patient which has the lowest sequence number within its respective admission. The set of all a patient's procedures is found by joining procedure and admission information through the admission ID. A patient has an event if they participate in

an admission, and that admission involves that procedure:

$$admitted(PatientID, AdmissionID, DateTime),$$
$$procedure(AdmissionID, ProcedureSeq, ProcedureDate, IcdCode, IcdVer) \quad (1)$$
$$\rightarrow event(PatientID, AdmissionID, ProcedureDate, ProcedureSeq, IcdCode, IcdVer)$$

Then, an aggregation function is used to introduce a new variable *minSeq*, the minimum *seq* value of procedures within that admission:

$$event(PatientID, AdmissionID_A, Date_A, Seq_A, IcdCode_A, IcdVer_A),$$
$$event(PatientID, AdmissionID_B, Date_B, Seq_B, IcdCode_B, IcdVer_B),$$
$$MinSeq = \min(Seq_B). \quad (2)$$
$$\rightarrow minSequence(PatientID, AdmissionID_A, Date_A, Seq_A, IcdCode_A, IcdVer_A, MinSeq)$$

Finally, a *primaryEvent* is defined as any event whose *Seq* is the lowest for that admission:

$$minSequence(PatientID, AdmissionID, Date, Seq, IcdCode, IcdVer, MinSeq),$$
$$Seq = MinSeq \quad (3)$$
$$\rightarrow primaryEvent(PatientID, AdmissionID, Date, IcdCode, IcdVer)$$

Frequently, we want to reduce a patient's entire record further, for example only including those events that are related to a particular disease of interest. Typically, this involves establishing a list of approved codes in consultation with subject matter experts — in practice, implemented by defining a relation *includedEvent*(X) that imports these events from a file. However, with the addition of domain information, this same task can be accomplished through reasoning. We use domain knowledge from the OMOP Common Data Model [24], which unifies several medical vocabularies, including ICD-9 Volume 3 and ICD-10-PCS's procedure codes (as used in the MIMIC-IV dataset), as well as SNOMED-CT, the most comprehensive available source of medical concepts. By integrating this additional source of knowledge, which encompasses several different sources itself, we can combine the benefits of each: using SNOMED-CT's rich semantics to discover concepts of interest, then automatically translating them to their equivalent codes in the observational data.

For example, if we are investigating colon cancer as our disease of interest, we might say as a first step that we only want to consider patient's procedures involving the colon. We build a list of concepts that represent the colon itself, or parts thereof: a concept is a *colonConcept* if it is a descendent of the concept "colon" (4215634). Firstly, we establish that *isA* is transitive:

$$isA(A, B), isA(B, C) \rightarrow isA(A, C) \quad (4)$$

$$isA(S1, S2), \ S2 = 4215634$$
$$\rightarrow colonConcept(S1) \quad (5)$$

Secondly, we define a *colonProcedure* as any procedure that has one of our discovered anatomy concepts marked as a *direct procedure site*:

$$dirProcSite(P, S), colonConcept(S) \rightarrow colonProcedure(P) \quad (6)$$

This returns a set of 1,097 procedure concepts, across the SNOMED, ICD-9 and ICD-10 systems. Thus, the process of identifying which codes to include and which to ignore in a model of a patient's pathway can be, in part, automated. For reliable and reproducible epidemiological studies, these will no doubt need to be audited by a domain expert, but the task of drawing up an initial codelist to iterate upon has been significantly shortened to only three rules.

As well as identifying appropriate events for exclusion and inclusion, it is also helpful to be able to abstract events into more general categories: knowing that a single patient experienced a *resection of sigmoid colon, percutaneous endoscopic approach* is less useful than knowing that a number of patients all underwent surgery for their colon cancer. As an example, and to provide a base dataset for later case studies, we use SNOMED's structure to retrieve all descendants of the concepts *operation on colon, chemotherapy, imaging of gastrointestinal tract* and *gastrointestinal system, inspection*, mapping these to the categories *Surgery, Chemotherapy, Imaging*, and *Colonoscopy* respectively. We create a new relation *eventCategory* that maps an ICD code to a text descriptor of its general category. For example, a code maps to the "chemotherapy" category if it is a descendent of the concept *chemotherapy*, with the reference to *concept(...)* required to map the concept's ID to its code in the ICD system.

$$isA(ConceptID_1, ConceptID_2), \ ConceptID_2 = 4221694,$$
$$concept(ConceptID_1, Code) \qquad (7)$$
$$\rightarrow \ eventCategory(Code, \text{``Chemotherapy''})$$

This rule is repeated for each of our four categories. We then define a new subset of events to focus our attention on: colon-specific events, or a simplified *event* in which the ICD code and version is replaced by a text description of its category.

$$primaryEvent(PatientID, AdmissionId, Date, IcdCode, IcdVer),$$
$$eventCategory(IcdCode, Desc) \qquad (8)$$
$$\rightarrow \ colonEvent(PatientID, AdmissionId, Date, Desc)$$

Further analysis can then be performed on this abstracted data, as we outline in the next section. The original data model, and the additional enhancements added by these rules, are illustrated in Figure 1. Our concept of an event is created by combining admissions and procedure observations with ontological knowledge, and then further definitions of events are layered on top of this representation.

This process is not perfect; one noticeable issue was that in the data model used, many ICD-9 chemotherapy concepts were not properly integrated and did not have *isA* relations to their relevant SNOMED concepts; a simple solution was to add an extra rule explicitly and also include the ICD-9 chemotherapy parent concept as a valid chemotherapy concept. This underlines the importance of choosing the right source of domain knowledge to match the given problem, and having a thorough understanding of its structure including limitations.

What counts as a "pathway" — and therefore which events are included or excluded — will be different for every research project. What is important is that each definition is explainable and auditable, with a clear set of inclusion and exclusion rules. Importantly, this approach again avoids the construction of a lengthy codelist. With a thorough and complete source of domain

knowledge existing, we only have to provide a relatively small number of constraints to narrow down our data to the events we need.

4.2. Cohort retrieval

When studying patient pathways, we often want to group patients according to their history. Such analysis makes it possible to identify patients by common treatment patterns, examine clusters, and detect outliers. For example, Tamm et al. group patients according to a simple sequence of events, defining cohorts such as "diagnosis → scan → surgery → scan" and "diagnosis → scan → chemoradiotherapy → radical resection → chemo(radio)therapy → scan" [25]. The ordering of the major stages in treatment can be an important research question with repercussions for care. One relevant example is the use of radiotherapy in treating rectal cancers. Because it is associated with *both* reduced recurrence and increased complications, judgements are often left to individual clinicians and local teams: consequently, usage dramatically varies across England [26]. Therefore, the ability to group patients by their sequences of events is a useful tool.

In order to extract a summary of each patient's history, we establish a new relation $next(A, B)$, meaning that two events A and B pertain to the same patient, and directly follow each other. More specifically, given any particular event A, we find all other events belonging to the same patient, and note their minimum date value (Rule 9):

$$
\begin{aligned}
colonEvent(&PatientID, EventID_A, Date_A, \_, \_, \_), \\
colonEvent(&PatientID, EventID_B, Date_B, \_, \_, \_), \\
&Date_A < Date_B, M = \min(Date_B) \\
\rightarrow\ &minDate(PatientID, EventID_A, Date_A, M)
\end{aligned}
\tag{9}
$$

A $next()$ relation exists between this original event, and the later event with the lowest date:

$$
\begin{aligned}
minDate(&PatientID, EventID_A, Date_A, Date_B) \\
colonEvent(&PatientID, EventID_A, Date_A, \_, \_, Desc_A), \\
colonEvent(&PatientID, EventID_B, Date_B, \_, \_, Desc_B) \\
\rightarrow\ &next(PatientID, EventID_A, EventID_B, Date_A, Date_B, Desc_A, Desc_B)
\end{aligned}
\tag{10}
$$

We also ensure that patients with only one event are still recorded, by creating an extra $next()$ relation that marks the end of a chain of events:

$$
\begin{aligned}
event(&P, A, Date_A, \_, \_, Desc_A), \\
&lastEvent(P, A, \_) \\
\rightarrow\ &next(P, A, 0, Date_A, 0, Desc_A, ".")
\end{aligned}
\tag{11}
$$

with *lastEvent*s being identified in a separate rule using the *max()* aggregation function, similar to Rule 9. This structure allows us to create a summary of each patient's events, through a combination of recursion and Vadalog's string operations. In the first case, a path of length

2 exists between any two events that follow each other with a *next()* relation; that path is summarised by concatenating the category descriptions of the two events.

$$next(PatientID, EventID_X, EventID_Y, \_, \_, Desc_X, Desc_Y),$$
$$String = \text{concat}(Desc_X, Desc_Y),$$
$$Length = 2 \tag{12}$$
$$\rightarrow path(PatientID, EventID_X, EventID_Y, String, Length)$$

In addition, a path exists between X and Z if there exists a next() relation between X and Y, and there already exists a path between Y and Z; this path's description is generated by concatenating X's onto the existing path's description.

$$path(PatientID, EventID_Y, EventID_Z, String, Length),$$
$$next(PatientID, EventID_X, EventID_Y, \_, \_, Desc_X, Desc_Y),$$
$$NewString = \text{concat}(Desc_X, String), NewLength = Length + 1 \tag{13}$$
$$\rightarrow path(PatientID, EventID_X, EventID_Z, NewString, NewLength)$$

Once the possible paths between adjacent events are recursively constructed, we simply find the longest path for each patient:

$$path(PatientID, EventID_X, EventID_Y, String, Len),$$
$$firstEvent(PatientID, EventID_X, \_),$$
$$path(PatientID, EventID_X, EventID_Z, String', Len'), \tag{14}$$
$$MaxLen = \text{mmax}(Length')$$
$$\rightarrow pathLength(PatientID, EventID_X, EventID_Z, String, Len, MaxLen)$$

$$pathLength(PatientID, EventID_X, EventID_Y, String, Len, MaxLen),$$
$$Len = MaxLen \tag{15}$$
$$\rightarrow longestPath(PatientID, EventID_X, EventID_Y, String, Len)$$

Finally, we can again apply an aggregation function to discover the most common paths followed by patients:

$$longestPath(PatientID, EventID_X, EventID_Y, String, Length),$$
$$Count = \text{mcount}(String) \tag{16}$$
$$\rightarrow pathFrequencies(String, Count)$$

These most common paths are shown in Table 1. The most common paths are unsurprisingly the simplest. As paths grow longer, they necessarily become more idiosyncratic, tailored to a specific case; in particular we observed that many of the longer paths contain repeated chemotherapy events, given that chemotherapy involves repeated cycles of treatment. For some analyses, it may be desirable to collapse these repeated events into one event that represents an entire treatment plan; this could be easily accomplished by incorporating an extra constraint $Desc_A \neq Desc_B$ into 9. This declarative approach to data access therefore simplifies not just *abstracting* events, but also *aggregating* them.

Table 1
The most common patient pathways

Path	Count
Surgery.	329
Chemotherapy.	18
SurgerySurgery.	18
Colonoscopy.	7
ChemotherapyChemotherapy.	5
[15 more]	<5

There are still some surprising results: few imaging procedures ever appear, possibly because these events are far more reliably recorded in the dataset's separate imaging table. Similarly, more representative pathways might be derived by retrieving chemotherapy information from the prescriptions table. However, this demonstration proves that such an approach to constructing pathways is possible and even intuitive to perform.

4.3. Guideline conflicts

Finally, in a third case study we demonstrate how additional data from outside of structured knowledge sources can complement observational data to produce new insights. The clinical guidelines that describe patient pathways are largely written for a single medical condition, but in reality patients often have overlapping multiple medical conditions (*comorbidities*). This means that patients are often prescribed multiple drugs at once, which can lead to a risk of dangerous side effects if those drugs cause adverse reactions. Dumbreck et al.'s 2015 study examines how often the guidelines for commonly co-occurring diseases recommend potentially dangerous drugs: choosing guidelines for three major conditions (heart failure, type 2 diabetes, and depression), they manually examine whether the recommended drugs conflict with the drugs for nine other common comorbidities [27]. The results indicated that potentially serious interactions were common. In this section, we reproduce the logic used in this study within Vadalog, demonstrating how such analysis, which required manually comparing and interpreting data from different sources, can be expressed logically in an intuitive and compact way.

The rules used to look for guideline conflicts can be fairly simply expressed: two guidelines are in potential conflict with each other if one recommends a drug D_1 that is known to interact with a drug D_2 recommended by the other.

$$recommends(G_1, D_1),\ recommends(G_2, D_2),$$
$$interacts(D_1, D_2) \tag{17}$$
$$\rightarrow conflict(G_1, G_2, D_1, D_2)$$

To obtain some data on guideline-recommended drugs, we used the MetaMap tool [28] to identify medical concepts in the text of guidelines from the UK's National Institute for Health and Care Excellence, considering any concept of semantic type "pharmacologic substance". In some cases, guidelines used in the original study were no longer maintained (for example, CG48 "secondary prevention for patients following myocardial infarction" has been merged into "acute coronary syndroms"): in these cases, the succeeding clinical guideline was used.

Table 2

Number of drug-drug interactions identified between heart disease and comorbidities

G_2 guideline	C conflicts
Atrial fibrillation	51
Depression in adults	45
Acute coronary syndromes	38
Chronic heart failure in adults	38
Dementia	30
Hypertension in adults	27
Chronic obstructive pulmonary disease in over 16s	26
Type 2 diabetes in adults	21
Chronic kidney disease	16
Neuropathic pain in adults	8
Rheumatoid arthritis in adults	3

Drug-drug interactions, meanwhile, were scraped from the online edition of the British National Formulary [29].

Once these datasets are in place, we first obtain a set of drugs that are specifically recommended for treating either heart failure, or one of the nine potential comorbidities considered:

$$
\begin{aligned}
recommends(G_1, D_1), \ G_1 &= \text{"\textit{Heart failure}"}, \\
recommends(G_2, D_2), \ potentialComorbidity&(G_2) \\
interacts&(D_1, D_2) \\
\rightarrow heartFailureConflict&(G_1, G_2, D_1, D_2)
\end{aligned}
\tag{18}
$$

Then, an aggregation function is used to count the number of conflicts.

$$
\begin{aligned}
heartFailureConflict(G_1, G_2, D_1, D_2), \\
C = \text{mcount}(G_2) \\
\rightarrow counts(G_2, C)
\end{aligned}
\tag{19}
$$

The results of this are shown in Table 2. Compared to Dumbreck et al.'s study as a baseline, these results bear a strong resemblance but also contain some noticeable differences. Arthritis, neuropathic pain, chronic obstructive pulmonary disease and chronic kidney disease still have the lowest number of conflicts with heart failure, and coronary syndromes, and atrial fibrillation still have the highest. More surprising results are that chronic heart failure, depression and dementia are more likely to conflict, whilst type two diabetes is far less likely. There are many possible reasons for these differences: changes in recommended drugs over time, changes in the scope of the guidelines, or the fact that many of the interactions were not considered to be serious enough by the previous study. These results inevitably require analysis and interpretation by domain experts to identify the specific mechanisms and reasons for changes, but the key point is that the initial data preparation and problem specification can now be completed in seconds, speeding up a laborious human process and in the future allowing such questions to be quickly reproduced across the whole spectrum of human disease.

5. Discussion and conclusions

In this paper, we introduced a key challenge in electronic health record data: the mapping, analysis, and interpretation of patient pathways. The pathway perspective on healthcare data is vital for understanding not just individual treatments, but the entire healthcare journeys of real-world patients, and the relationship between official guidelines and true clinical practice. We highlight the benefits of using a declarative language for these tasks, in particular the ability to allow complex relationships between many entities to be expressed as intuitive rules, paving the way for domain experts to act as a data analysts by expressing their own rules and definitions. In the three case studies presented, we considered how a reasoning language such as Vadalog can be applied to real-world data, ontological knowledge, and domain free-text to generate new insights into patient care, and how several of the language's features such as recursion and aggregation functions can simplify tasks. Given that so much data-driven research is currently based on proprietary scripts developed by individual researchers, leading to enormous duplication of effort, we believe declarative approaches to be a key component in moving towards a more transparent and reproducible approach to health data science. Exploring health data in terms of declarative rules allows meaning to be constructed from raw data in an explainable way. We saw for example in our first case study how a healthcare event can be reinterpreted as a "primary event" and then as a "colon event", at each stage the user adding their own layers of meaning and interpretation that fit the needs of their own research.

This paper aims to introduce this novel domain – cancer pathway knowledge graphs – as an area of potential interest for the rules and reasoning community, and provide several proofs of concept on publicly available data that provide a solid foundation for answering interesting healthcare questions. However, there still exists an abundance of open challenges requiring further attention. Whilst most of the data discussed here is readily provided in a structured format, we also touch on free-text data in our third case study because there remains much EHR information stored as free text. In particular, cancer staging and recurrence is often hidden in imaging or pathology reports, requiring natural language processing approaches to prepare them into a form suitable for logic-based reasoning [25]. The handling of this data, and its conversion into suitable formats, is still an open research question.

We also envisage that future research might consider more complex pathway related questions, for example matching patients to the most similar guideline based on their pathway, or evaluating recommended pathways in terms of how closely real-world patients follow them. Sub-symbolic machine learning methods such as clustering, classification, and knowledge graph embedding offer promise in this regard. However they naturally raise issues of transparency, reproducibility, and bias. An important way forward in the future will likely be the integration of these methods for advanced pattern recognition and discovery with rigorous logical formulations that align them with formal domain knowledge, allowing the best of both approaches to be exploited.

In summary, we believe the challenge of clinical pathway knowledge graphs to be both a currently important societal challenge, and an important challenge for our declarative AI community that is particularly accessible to declarative methods. While medical knowledge graphs and ontologies have long been a staple of our field, we believe that the current push in large national and international projects for cancer pathway research is a great chance for the declarative AI community to contribute to a fresh and challenging topic.

Acknowledgments

References

[1] T. Alfonsi, B. Luigi, A. Bernasconi, S. Ceri, Expressing Biological Problems with Logical Reasoning Languages, in: Rule Challenge at RuleML+RR 2022, volume 3229 of *CEUR Workshop Proceedings*, CEUR-WS, 2022.

[2] L. Bellomarini, D. Benedetto, G. Gottlob, E. Sallinger, Vadalog: A modern architecture for automated reasoning with large knowledge graphs, Inf. Syst. 105 (2022) 101528.

[3] P. Atzeni, L. Bellomarini, M. Iezzi, E. Sallinger, A. Vlad, Augmenting logic-based knowledge graphs: The case of company graphs, in: KR4L at ECAI 2020, volume 3020 of *CEUR Workshop Proceedings*, CEUR-WS, 2020, pp. 22–27.

[4] P. Atzeni, L. Bellomarini, M. Iezzi, E. Sallinger, A. Vlad, Weaving enterprise knowledge graphs: The case of company ownership graphs, in: EDBT 2020, OpenProceedings.org, 2020, pp. 555–566.

[5] L. Bellomarini, L. Bencivelli, C. Biancotti, L. Blasi, F. P. Conteduca, A. Gentili, R. Laurendi, D. Magnanimi, M. S. Zangrandi, F. Tonelli, S. Ceri, D. Benedetto, M. Nissl, E. Sallinger, Reasoning on company takeovers: From tactic to strategy, Data Knowl. Eng. 141 (2022) 102073.

[6] T. Baldazzi, D. Benedetto, M. Brandetti, A. Vlad, L. Bellomarini, Heuristic-based reasoning on financial knowledge graphs, in: EDBT/ICDT Workshops, volume 3135 of *CEUR Workshop Proceedings*, CEUR-WS, 2022.

[7] T. Baldazzi, D. Benedetto, M. Brandetti, A. Vlad, L. Bellomarini, E. Sallinger, Datalog-based reasoning with heuristics over knowledge graphs, in: Datalog 2.0 at LPNMR, volume 3203 of *CEUR Workshop Proceedings*, CEUR-WS, 2022, pp. 114–126.

[8] A. Vlad, S. Vahdati, M. Nayyeri, L. Bellomarini, E. Sallinger, Towards hybrid logic-based and embedding-based reasoning on financial knowledge graphs, in: EDBT/ICDT Workshops 2022, volume 3135 of *CEUR Workshop Proceedings*, CEUR-WS, 2022.

[9] T. Rotter, L. Kinsman, E. James, A. Machotta, H. Gothe, J. Willis, P. Snow, J. Kugler, Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs, Cochrane Database of Systematic Reviews (2010) CD006632.

[10] D. Allen, E. Gillen, L. Rixson, Systematic review of the effectiveness of integrated care pathways: what works, for whom, in which circumstances?, Int. J. Evid. Based Healthc. 7 (2009) 61–74.

[11] J. A. M. Stothers, A. Nguyen, Can Neo4j Replace PostgreSQL in Healthcare?, AMIA Jt. Summits Transl. Sci. Proc. 2020 (2020) 646–653.

[12] B.-H. Yoon, S.-K. Kim, S.-Y. Kim, Use of Graph Database for the Integration of Heterogeneous Biological Data, Genom. Inform. 15 (2017) 19–27.

[13] W. S. Campbell, J. Pedersen, J. C. McClay, P. Rao, D. Bastola, J. R. Campbell, An alternative database approach for management of SNOMED CT and improved patient data queries, J. Biomed. Inform. 57 (2015) 350–357.

[14] R. Piro, Y. Nenov, B. Motik, I. Horrocks, P. Hendler, S. Kimberly, M. Rossman, Semantic Technologies for Data Analysis in Health Care, in: ISWC 2016, LNCS, Springer, 2016, pp. 400–417.

[15] T. Oliveira, P. Novais, J. Neves, Development and implementation of clinical guidelines: An artificial intelligence perspective, Artif. Intell. Rev. 42 (2014) 999–1027.

[16] International Organization for Standardization, System of concepts to support continuity of care (ISO 13940:2015), 2015. URL: https://iso.org/standard/58102.html.

[17] L. Bellomarini, G. Gottlob, A. Pieris, E. Sallinger, Swift logic for big data and knowledge graphs, in: IJCAI, Springer, 2017, pp. 2–10.

[18] A. Calì, G. Gottlob, T. Lukasiewicz, A general Datalog-based framework for tractable query answering over ontologies, J. Web Semant. 14 (2012) 57–83.

[19] C. Zaniolo, M. Yang, A. Das, A. Shkapsky, T. Condie, M. Interlandi, Fixpoint semantics and optimization of recursive Datalog programs with aggregates, Theory Pract. Log. Program. 17 (2017) 1048–1065.

[20] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. Celi, R. Mark, MIMIC-IV (ver 1.0), 2020.

[21] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, R. G. Mark, MIMIC-IV, a freely accessible electronic health record dataset, Sci. Data 10 (2023) 1.

[22] MIT Laboratory for Computational Physiology, MIMIC-IV documentation: procedures_icd, accessed 2023-07-12. URL: https://mimic.mit.edu/docs/iv/modules/hosp/procedures_icd/.

[23] National Clinical Coding Standards OPCS-4, 9.1 ed., National Health Service (UK), 2022. URL: https://classbrowser.nhs.uk/ref_books/OPCS-4.9_NCCS-2022.pdf.

[24] OMOP Common Data Model v5.4, accessed 2023. URL: http://ohdsi.github.io/CommonDataModel/cdm54.html#Vocabulary_Tables.

[25] A. Tamm, H. J. Jones, W. Perry, D. Campbell, R. Carten, J. Davies, et al., Establishing a colorectal cancer research database from routinely collected health data: the process and potential from a pilot study, BMJ Health Care Inform. 29 (2022) e100535.

[26] E. J. A. Morris, P. J. Finan, K. Spencer, I. Geh, A. Crellin, P. Quirke, J. D. Thomas, S. Lawton, R. Adams, D. Sebag-Montefiore, Wide Variation in the Use of Radiotherapy in the Management of Surgically Treated Rectal Cancer Across the English National Health Service, Clin. Oncol. 28 (2016) 522–531.

[27] S. Dumbreck, A. Flynn, M. Nairn, M. Wilson, S. Treweek, S. W. Mercer, P. Alderson, A. Thompson, K. Payne, B. Guthrie, Drug-disease and drug-drug interactions: systematic examination of recommendations in 12 UK national clinical guidelines, BMJ 350 (2015) h949.

[28] A. R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, Proc. AMIA Symp. (2001) 17–21.

[29] British National Formulary, 2023. URL: https://bnf.nice.org.uk/.