COPE: Chronic Observation and Progression Events Ontology

Asara Senaratne[1,*], Oshani Seneviratne[2], Hon Zent Lim[1] and Leelanga Seneviratne[3]

[1]College of Science and Engineering, Flinders University, South Australia.
[2]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA.
[3]Faculty of Information Technology, University of Moratuwa, Sri Lanka.

Abstract

Artificial Intelligence (AI) is increasingly applied in healthcare to enable early detection, personalized prediction, and proactive management of chronic diseases. However, their effective integration remains hindered by the lack of a semantic framework that consolidates evolving patient trajectories, intervention pathways, and computational methods. This gap leads to significant challenges in model interpretability, reproducibility of research findings, and the systematic discovery of actionable insights from longitudinal patient data. In this work, we present the Chronic Observation and Progression Events (COPE) Ontology, an ontology designed to support health informaticians and data scientists in modelling chronic disease progression and aligning it with AI-driven approaches. Using the COPE ontology, we can capture patient related knowledge, including biological, behavioural, demographic, psychographic, and geographical characteristics, as well as modifiable and non-modifiable risk factors, symptom progression, and disease outcomes. The COPE ontology introduces a temporal structure that enables the representation of timestamped clinical events, symptoms, interventions, and exposures, thereby facilitating detailed modelling of disease trajectories over time. Also, COPE integrates AI/ML techniques for analyzing chronic disease conditions gleaned from scholarly literature. It further embeds the provenance of computational models, linking them to the six datasets the models use, the twenty five research literature from which they originate, and the specific health contexts in which they are applied. By bridging patient characteristics, temporal health trajectories, intervention strategies, and AI/ML capabilities within a unified semantic framework, the ontology provides a robust foundation for interpretable, reproducible, and patient oriented decision support. We demonstrate its utility through exemplar queries, offering it as a reusable resource for advancing the integration of AI in health trajectory modelling for chronic disease care.

Keywords

Health Trajectory Modelling, Disease Progression, Chronic Diseases, Temporal Events, Artificial Intelligence

1. Introduction

Chronic diseases such as diabetes, cardiovascular conditions, and respiratory illnesses represent a significant burden on global healthcare systems, accounting for nearly 74% of all deaths worldwide [1]. Early detection and effective intervention are critical to managing these conditions and improving patient outcomes. In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as powerful tools in this space, enabling prediction, stratification, and personalized care planning [2, 3].

A key challenge in advancing AI-driven health systems for chronic disease management lies in the absence of a semantic framework that cohesively brings together knowledge on patient disease trajectories, potential interventions over time, and the computational models used to support decision-making. Traditional approaches often create fragmented knowledge [4], making it challenging to understand the origin of AI model predictions, compare intervention effectiveness across different patient groups, or identify new disease progression patterns systematically [5]. This fragmentation hinders the creation of interpretable, reproducible, and patient-centered systems, thereby limiting our ability to gain actionable insights for precision medicine. Although notable ontology-driven explainable approaches have emerged for chronic diseases in clinical settings [6, 7, 8], a cohesive

5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, Nov 2025, Nara, Japan
*Corresponding author.
asara.senaratne@flinders.edu.au (A. Senaratne); senevo@rpi.edu (O. Seneviratne); lim0862@flinders.edu.au (H. Z. Lim); leelangas@uom.lk (L. Seneviratne)

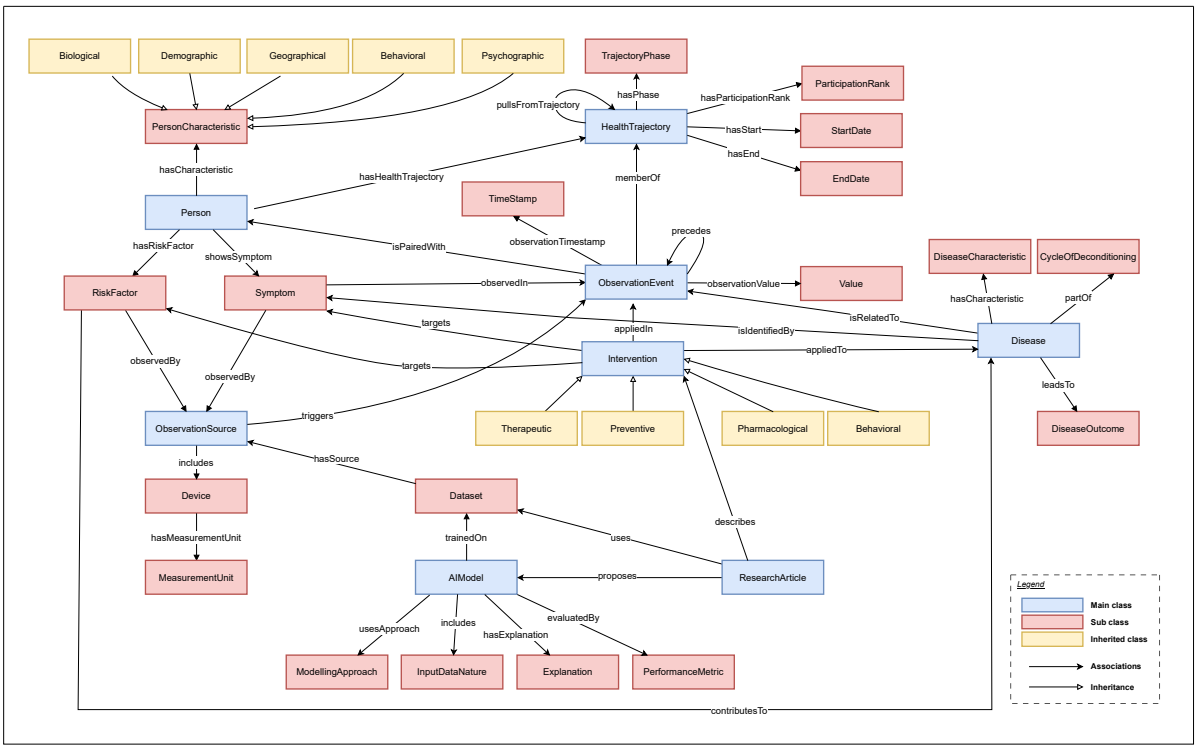

Figure 1: Conceptual model of the Chronic Observation and Progression Events (COPE) Ontology.

solution that integrates various aspects for developing comprehensive patient trajectories is still needed. An ontology, a formal and structured representation of knowledge [9], offers a powerful means to address this challenge by capturing complex interdependencies between patient characteristics, disease progression, interventions, and computational analysis in a machine interpretable manner [10].

To address this gap, we propose Chronic Observation and Progression Events (COPE), an ontology that unifies knowledge around the modelling of health trajectories in chronic disease contexts, the timing and nature of interventions across these trajectories, and the AI/ML methods used to derive clinical insights. We visualize the conceptual model of COPE in Figure 1. Unlike existing ontologies [11, 12, 13], which often focus either only on biomedical entities or algorithmic taxonomies, our approach holistically integrates biological, behavioural, psychographic, demographic, and geographical patient characteristics with modifiable and non-modifiable risk factors, symptom evolution, and disease outcomes.

The COPE ontology includes classes and properties that formalize relationships between patient profiles, observed symptoms, clinical interventions, and the dynamic progression of chronic illnesses. It also semantically links these entities to computational models, capturing details of AI/ML techniques (such as classification, regression, and clustering), data types (such as structured, temporal, and multi-modal), sources (such as EHRs, ECG, and wearable sensors), and scholarly outputs (such as datasets, publications, and venues).

A major contribution of the ontology is its support for temporal reasoning. Through the *Observation-Event* class and its associated properties, it models timestamped clinical observations and intervention records, enabling representation of time-dependent health trajectories (we describe in Section 3). This temporal scaffolding allows for complex, longitudinal queries such as; *What symptoms emerged after a specific intervention?* or *How did the patient's risk profile evolve across disease stages and episodes?*

We also capture AI/ML models and datasets used in the literature for health trajectory modelling. The purpose is to support health informaticians and data scientists by bringing this fragmented knowledge into one structured, accessible space. By doing so, the ontology enables users to identify and reuse established modelling techniques, replicate prior studies, compare performance across datasets, and build upon proven approaches. It facilitates systematic exploration of past research, helping practitioners

avoid redundant efforts while ensuring alignment with clinically validated practices. Ultimately, this promotes more efficient, evidence-based development of technical solutions tailored to chronic disease trajectories.

Overall, the COPE ontology is designed to serve three interconnected purposes: (1) to enable semantic modelling of chronic disease trajectories using episodic or continuous data from diverse sources such as wearable devices and clinical systems, (2) to support health informaticians, behavioral scientists, and data scientists in systematically mapping AI/ML contributions to chronic disease prediction and management, thereby accelerating scientific understanding and evidence-based practice by providing a semantically rich foundation for advanced analytical methods, and (3) to predict suitable interventions based on disease and patient profiles.

We demonstrate the utility of COPE through a set of competency questions and example queries that highlight its application in chronic disease research and decision support. In Section 2, we present a review of existing literature in the domains of designing ontology and health trajectory modelling. We present our COPE ontology in Section 4 and its evaluation in Section 5. We conclude the paper setting the future directions in Section 6.

2. Literature Review

In recent years, the transformation of healthcare through data has spurred the need for intelligent, interoperable systems that can represent, reason about, and act upon complex medical knowledge. Traditional electronic health records (EHRs) and siloed datasets, while rich in information, often fail to capture the nuanced interplay between clinical factors such as patient characteristics, disease progression, symptom evolution, risk factors, devices used for symptoms measuring, and therapeutic interventions. This gap has motivated researchers and clinicians alike to turn toward ontologies for formal and machine readable representations of domain knowledge as a means to integrate and analyse such multidimensional information more effectively.

Foundational biomedical ontologies like Systematized Nomenclature of Medicine: Clinical Terms (SNOMED CT) [11], Logical Observation Identifiers Names and Codes (LOINC) [12], and the Gene Ontology [13] have laid the groundwork by establishing standardized vocabularies for clinical observations, laboratory tests, and genetic factors. These efforts enable interoperability at a syntactic level but often lack the semantic expressiveness needed for modelling when a symptom occurs, how a disease evolves, or which risk factors precede a particular outcome. While the Human Disease Ontology [14] and Human Phenotype Ontology [15] provide well curated vocabularies for classifying diseases and phenotypic features, they do not offer native support for representing temporal sequences or linking these concepts to evidence from real-world datasets and literature.

2.1. Temporal Clinical Ontologies & Event-Based Modelling

Time is a fundamental aspect of clinical reasoning. Diseases unfold over time, interventions are applied at specific moments, and symptoms often emerge in unpredictable patterns. Despite these, most existing healthcare ontologies are atemporal. They represent what exists, but not when it exists. The W3C Time Ontology in OWL [16] provides primitives such as instant, interval, and duration, which are designed to model temporal relationships. However, their adoption in healthcare ontologies has been limited. One effective workaround proposed in ontology engineering is event introduction. That is, modelling events such as *a symptom was observed* or *a treatment was applied* as first class entities, which allows temporal and contextual metadata to be attached to clinical interactions. More recent efforts, such as the Time Event Ontology (TEO) [17], explicitly model complex temporal relationships found in clinical narratives, demonstrating advanced capabilities for representing temporal reasoning in healthcare. Furthermore, frameworks like OCEP [18] utilize ontology-based complex event processing for healthcare decision support, integrating real-time data from various sources to identify dynamic clinical patterns.

2.2. Semantic Integration of AI/ML Models with Clinical Data

Parallel to this development, the growing use of AI and ML in clinical decision support systems (CDSS) introduces a new modelling challenge [19]. Although AI/ML models are increasingly relied upon to predict disease onset, identify risk factors, and recommend interventions, they are often treated as black boxes with interpretation difficulties and poorly integrated into semantic healthcare frameworks. Ontologies such as ML-Schema [20] and Ontology of Core Data Mining Entities (OntoDM-Core) [21] have emerged to address this gap by describing ML workflows and metadata. However, their use remains largely separate from healthcare ontologies and does not yet enable the modelling of AI models in conjunction with specific patient data, symptoms, or interventions [22, 23].

Furthermore, there is a notable disconnect between scientific evidence and its formal representation in clinical ontologies. Ontologies such as the Bibliographic Ontology (BIBO) [24] and the Semantic Publishing and Referencing (SPAR) Ontology [25] enable the formalization of metadata about research publications, yet they are rarely linked to the clinical phenomena they investigate. In evidence-based medicine, however, it is vital to trace findings back to the articles, authors, and venues that reported them, specially when these findings underpin AI models used in real-world healthcare decisions.

2.3. Semantic Clinical Pathway/Guideline Ontologies

Health trajectory modelling and prevention strategy frameworks, such as the American Heart Association (AHA) taxonomy for chronic disease management [26], aim to categorize how diseases progress over time and how interventions can alter these trajectories. Some taxonomies explicitly incorporate trajectory or disease stages. For instance, life-course frameworks and traditional public health models categorize interventions by when they occur relative to disease progression (such as pre-disease, early disease, established disease, and so on) [27]. Such taxonomies are valuable for organizing knowledge across diverse diseases and populations, guiding research and clinical decision-making. However, existing taxonomies vary widely in structure and focus. A narrowly scoped taxonomy can be deeply detailed for that domain but may not extend easily to other conditions. Conversely, broad taxonomies often categorize by disease type (such as communicable, non-communicable, mental health, and injury) and by intervention strategy, to ensure all health issues are covered. A scalable taxonomy should ideally allow cross-cutting modules, for example, one axis for disease category and another for intervention type or data type, so that it can expand as new diseases or strategies are considered. Beyond mere categorization, there is a growing need for semantic representation of clinical pathways and guidelines to ensure standardized, evidence-based care. Ontologies like ShaRE-CP (Shareable and Reusable Clinical Pathway) Ontology [28] focus on formalizing clinical pathways, including temporal constraints between interventions and linking them to relevant health resources and contextual information. Such efforts are crucial for enabling automated reasoning about treatment plans, identifying optimal intervention sequences, and ensuring adherence to best practices, which aligns with COPE's goal of supporting intervention prediction based on patient and disease profiles.

In summary, while existing ontologies have made strides in formalizing healthcare knowledge, most remain limited in temporal scope, disconnected from AI workflows, and loosely coupled to scientific evidence. Therefore, the ontology introduced in this paper aims to fill these gaps by creating a unified, extensible, and temporally grounded ontology that reflects both clinical processes and computational intelligence, aligned with the needs of modern health systems and data science applications.

3. Health Trajectory

Based on the emerging Interpretive Paradigm for knowledge representation [4] and sensitive data representation [30], we define a health trajectory as a collection of sequentially linked observer events that share the same membership (trajectory membership). We visualize this in Figure 2. The health trajectory is local to a specific entity, such as a person or a community. Contrary to existing trajectory modelling, our approach detects and allows multiple normalities to exist simultaneously. At any given

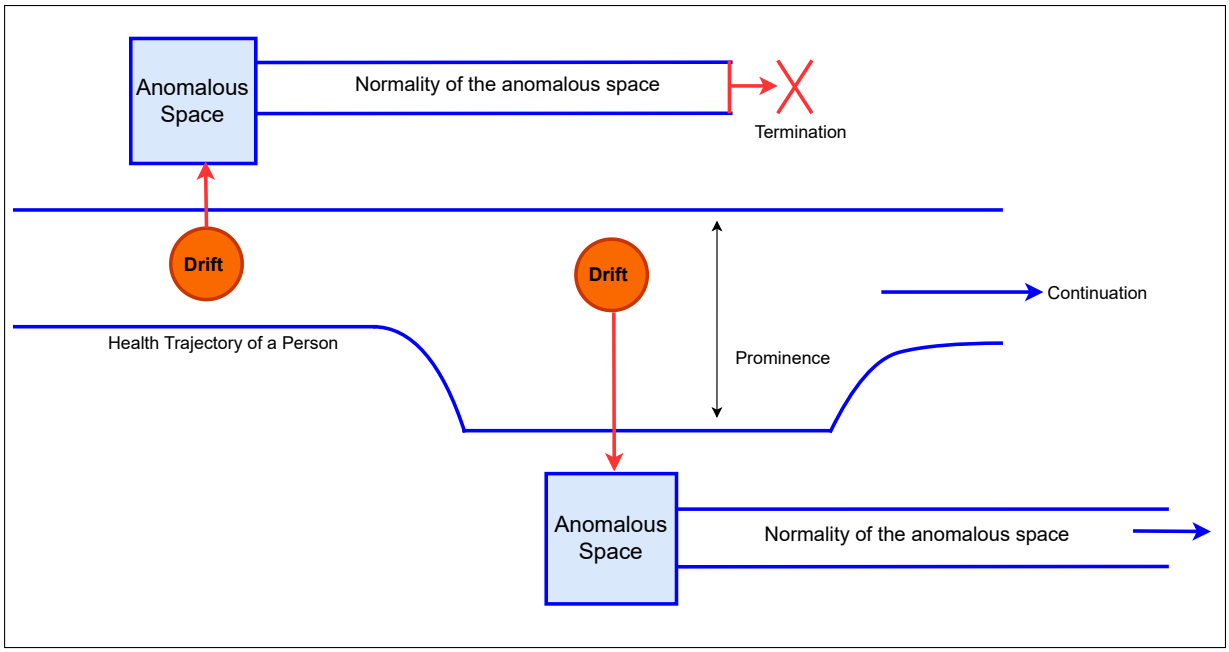

Figure 2: Conceptual overview of a health trajectory as we consider in COPE [29].

temporal point, the entity (a person in this case) accommodates one trajectory out of several possible others. These alternative trajectories are modelled in anomalous spaces [31, 32, 33] relative to the current trajectory the person traverses. The trajectory a person is in, is called the participating trajectory, which will be denoted by a participation rank of 0. The ranks of the other trajectories of the same person indicate the relative pull effect to become the participating trajectory.

When trajectory progression approaches a defined drift point, the propensity to deviate is governed by two principal forces: the pull effect and the withdrawal effect. The pull effect is quantified by the convergence index, which measures the degree of alignment between the current progression and the set of all feasible trajectories at time t_n. In contrast, the withdrawal effect is determined by the latching strength of the present trajectory, expressed as the conditional probability that the progression will remain within the current trajectory versus the likelihood of disengagement and transition into an alternative trajectory.

The tendency of a person to progress through the same trajectory, given the circumstances, is represented as prominence. A drift point is a moment when an individual's health trajectory changes significantly, potentially pulling them from one trajectory to another. For example, from a healthy state to a deteriorating one. Accordingly, health trajectory modelling involves the assessment of sequentially linked observer events as a sequence of states over time. Each state will correspond to a particular combination of physiological, emotional, and behavioural data. Machine learning techniques such as LSTM (Long Short-Term Memory) networks potentially model these sequences, capturing the relationships between different states and how they evolve. Here we study how individuals transition from one trajectory to another by studying drift points over time. For example, we examine how someone moves from a *stable* to a *declining* health trajectory [29].

3.1. Use Case

Having defined the conceptual structure of health trajectories, we now ground these ideas in a concrete use case. Consider a diabetes management scenario. A patient's health trajectory includes timestamped laboratory observations (HbA1c values, BMI), wearable-derived data (daily step counts, heart rate variability), and behavioural attributes (dietary habits, stress levels). These are represented in COPE as `ObservationEvents` linked to `PersonCharacteristics` and `RiskFactors`.

Trajectory progression. At baseline, the patient resides in a *stable* `TrajectoryPhase`. Over time,

consistently elevated HbA1c values and sedentary behaviour introduce a drift point, triggering a transition into a *deteriorating* trajectory.

Interventions. Pharmacological treatments (e.g., metformin) and behavioural programs (e.g., increased exercise) are modelled as instances of the `Intervention` class, applied to specific `TrajectoryPhases`. Their effects can be queried across patients with similar risk profiles.

Outcome. This scenario demonstrates how COPE integrates patient data, temporal trajectories, and AI provenance into a coherent framework. Researchers can ask: Which interventions were most effective for patients with similar risk profiles? Which AI/ML models (e.g., LSTMs) have been applied to predict progression? While diabetes illustrates the workflow, COPE is designed to generalize across chronic disease contexts.

4. Ontology Overview

This section presents our COPE ontology, that was developed to formally represent the intersection between patient profile, disease profile, symptom dynamics, risk factors, and the integration of AI/ML methods for chronic disease modelling, as shown in Figure 1. We designed the ontology to support knowledge-driven reasoning and semantic querying in health informatics applications, particularly those involving time-aware patient trajectory analysis and explainable ML.

We developed the ontology with several key objectives in mind. First, it aims to support a patient-centered approach to disease modelling by capturing biological, psychological, behavioural, and demographic characteristics. Second, it is designed to represent the dynamic nature of chronic disease progression through temporally anchored clinical events and distinct phases of health trajectories. Third, the ontology incorporates AI and ML models that have been used in the literature for chronic disease management, prediction and health trajectory profiling. Fourth, it integrates research knowledge artifacts such as scientific publications, datasets, and evidence of model performance to support informed decision-making. Finally, the ontology enables reasoning and semantic querying across patient profiles, model metadata, and scholarly outputs, facilitating more intelligent exploration and reuse of knowledge in chronic disease research. The COPE ontology includes the following high-level classes:

- Person and Characteristics: The central class `Person` is associated with instances of `PersonCharacteristic`, categorized into subtypes such as `Biological`, `behavioural`, `Demographic`, `Psychographic`, and `Geographical`. These characteristics inform disease susceptibility and model stratification. Individuals are also linked to one or more `RiskFactor` instances.
- Symptoms and Observations: Clinical manifestations are represented by the `Symptom` class. Symptoms are temporally recorded via `ObservationEvent`, which encapsulates time-stamped symptom occurrence, observation source (such as device and Electronic Health Records: EHR), and optional quantitative values. This enables modelling symptom timelines and episodic patterns.
- Health Trajectories: Temporal sequences of observation events are grouped into a `HealthTrajectory`, which consists of ordered `TrajectoryPhase` instances (such as onset, exacerbation, and remission). Each phase is annotated with start and end dates, enabling longitudinal modelling of disease progression.
- Diseases and Interventions: Diseases are modelled via the `Disease` class and are connected to `Symptom`, `RiskFactor`, `CycleOfDeconditioning`, and `DiseaseOutcome`. Interventions are instances of the `Intervention` class, categorized into `Pharmacological`, `Therapeutic`, `Preventive`, and `behavioural`, and linked to either symptoms or diseases through `appliedTo` or `appliedIn`.
- Artificial Intelligence Models: AI models are represented using the `AIModel` class, associated with `Dataset` (via `trainedOn`), `ModellingApproach`, `InputDataNature`, and `Explanation`. Performance metrics are modeled using `PerformanceMetric` and linked to the models they evaluate.

Each model or dataset may be described or proposed in a `ResearchArticle`, connected via properties such as `describes` or `evaluates`. Articles are in turn linked to `Author`, `DOI`, and `Venue` entities.

4.1. Design and Development

We developed COPE following a structured approach, beginning with the creation of a taxonomy informed by requirements elicitation and literature analysis. To gather requirements related to personal characteristics and health trajectories, we consulted Leelanga Seneviratne, a behavioural scientist and a co-author of this paper. We also reviewed twenty five research papers focused on the application of AI in chronic disease prediction and management. Based on this input, we formulated competency questions to define the scope of the ontology, drawing on both our domain expertise and insights from the literature. During the conceptualization phase, we decomposed the domain into modular components, including patient characteristics, health trajectories, symptoms, observation events, risk factors, disease outcomes, clinical interventions, AI models, and research artifacts. In the formalization phase, we implemented these components in OWL 2 using Protégé, defining appropriate classes, object and data properties, domain and range axioms, disjointness constraints, and human-readable annotations such as `rdfs:label` and `rdfs:comment`. Finally, we evaluated the ontology by executing SPARQL queries based on our competency questions over RDF instances to ensure completeness, correctness, and alignment with the intended use cases.

To ensure interoperability with scholarly metadata standards and integration with Research Knowledge Graphs (RKGs), we reused elements from the **Scholarly Communication Ontology (SCIO)**. Specifically, we adopted:

- `scio:Person` represents an individual whose characteristics, behaviors, health trajectory, or related data are captured in the ontology.
- `scio:Disease` captures a medical condition or diagnosis associated with a person, including links to symptoms, risk factors, and outcomes.
- `scio:Device` refers to a data collection device that collects information about one or more objects.
- `scio:Intervention` Refers to a clinical action or treatment.
- `scio:Dataset` denotes a structured collection of data generated or used in research, enabling traceability, reuse, and citation through metadata.
- `scio:hasSource` a relation between an entity and another entity from which it stems from.
- `scio:describes` a relation between one entity and another entity that it provides a description of (detailed account).
- `scio:precedes` expresses a temporal or logical ordering between two entities, such as sequential interventions, observation events or stages in a process.
- `scio:isMemberOf` a mereological relation between a item and a collection.
- `scio:isPartOf` a transitive, reflexive and anti-symmetric mereological relation between a whole and itself or a part and its whole.

4.2. Release and Sustainability Strategy

To ensure best practices in ontology publication, COPE is assigned the persistent namespace: https://purl.archive.org/cope#, which is dereferenceable and ensures long-term accessibility. We have released the first public version (V1) of the ontology, which is available through the open-source repository on GitHub available at https://github.com/hzent/COPE. In addition to an interactive dashboard available at https://hzent.github.io/COPE/CopeDashboard.html, we provide a standardized WIDOCO [34]-generated documentation that describes the ontology's scope, comprehensive documentation of the classes and properties introduced, which is available from the ontology namespace https://purl.archive.org/cope. This documentation complements this paper and the dashboard, making it easier for others to adopt

Table 1
Competency questions and candidate answers grouped based on the aspect they cover in the ontology.

Question Categories	Competency Questions	Candidate Answers
Patient and Disease Associations	Which patient characteristics are most strongly associated with increased susceptibility to specific diseases?	A person who undergoes a high level of stress is more susceptible for anxiety.
	Which population groups are most susceptible to disease X based on their characteristics?	A person whose hobby is mountain hiking is more susceptible to Monge's disease.
	Given certain patient characteristics, what disease(s) are they at risk of developing in the near future?	A female having BRCA1/2 gene mutation is having high risk of getting breast cancer.
Risk and Disease Progression	What are the risk factors associated with lung cancer?	Smoking and prolonged exposure to Radon are risk factors associated with cancer.
	What diseases are patients at risk of, and what are the associated disease outcomes, and identifying symptoms?	Alcoholic person is more likely to get Hypertension and might be having risk of kidney failure.
Progression of Health Trajectories	Which trajectory does the person participate in at a specific observation event?	As per the observation event noted on the 3rd of August, HT01002 is the participating trajectory of the person.
	Which disease are the symptoms observed in within the current participating trajectory?	Fatigue and chest pain observed on the 3rd of August in the participating trajectory are related to Coronary Artery Disease.
AI/ML Techniques and Applications	Which AI/ML techniques are most commonly used for modelling disease progression?	LSTM and Transformers are commonly used for identifying the progression of stroke.
	Which research articles focus on the detection or prediction of chronic diseases using AI/ML methods?	Priyanga, P., Pattankar, V. V., & Sridevi, S. (2021). A hybrid recurrent neural network - logistic chaos-based whale optimization framework for heart disease prediction with EHRs. Computational Intelligence., 37(1), 315–343. https://doi.org/10.1111/coin.12405 is an article that proposes a hybrid RNN for predicting heart disease.
	What is the predominant nature of datasets used for training AI/ML models in disease prediction?	Unstructured clinical notes are used with text embeddings + SVM for determining Alzheimer's disease.
	What ML models and input features are commonly used for predicting diseases?	LSTM is commonly used with HbA1c and Sleep Patterns to assess the likelihood of having Diabetes Mellitus.
	For a given data source, which AI/ML techniques are most frequently applied?	For data from wearable devices, Random Forest is commonly used to recognize activities such as walking and climbing.

and extend COPE. Although COPE is at a prototype stage, we have outlined a clear strategy for long-term sustainability. Future releases will follow semantic versioning, with major and minor updates archived with persistent identifiers via purl to ensure reproducibility and traceability. Finally, to promote community engagement and reuse, the GitHub repository includes an issue tracker and will include contribution guidelines in upcoming releases. This will allow external contributors to propose new terms, raise issues, and suggest extensions.

4.3. Competency Questions

To evaluate the expressiveness and utility of the COPE ontology, we formulated a set of competency questions (as shown in Table 1) grouped across three core dimensions: Patient and Disease Associations, Risk and Disease Progression, and AI/ML Techniques and Applications. These questions serve as practical benchmarks to assess whether the ontology can adequately represent and retrieve knowledge relevant to chronic disease and trajectory progression, AI integration, and patient context.

The competency questions were derived from multiple sources. First, they were informed through discussions with domain experts, including a neuroinformatician, a digital health researcher, and a medical devices researcher, ensuring that the questions reflect practical considerations from medical research and practice. Second, we reviewed relevant literature in neuroinformatics and digital health to identify commonly addressed challenges and information needs. Together, these inputs provided both empirical grounding and theoretical justification, ensuring that the competency questions are not only intuitively valid but also aligned with practices and requirements observed in the medical domain.

The first set of questions, *Patient and Disease Associations*, focuses on uncovering relationships between individual characteristics and disease susceptibility. These questions evaluate whether the ontology can support queries such as: which patient attributes (such as stress levels, lifestyle, or genetic markers) are linked to heightened disease risk? and which demographic or behavioural profiles correspond to specific chronic conditions? For example, given a patient with the gene mutation, the ontology should allow retrieval of their elevated risk for breast cancer. These questions validate the ontology's ability to semantically model and reason over personal, biological, and environmental characteristics in relation to diseases.

The second set of questions, *Risk and Disease Progression*, examines how the ontology captures risk factors, symptom patterns, and expected outcomes. Questions in this category assess whether the ontology can express complex relationships, such as the link between alcohol consumption and hypertension, or between smoking and lung cancer. Moreover, they test whether the ontology can be queried to reveal probable disease outcomes based on combined risk profiles and symptom trajectories, thereby supporting use cases like early risk assessment and proactive intervention planning.

The third set of questions, *Progression of Trajectories*, addresses the temporal aspects of health trajectories, focusing on how patient conditions develop and unfold over time. This group captures questions that investigate the sequential and evolutionary nature of trajectories, such as identifying which trajectory follows another, how one state transitions into the next, or how multiple trajectories interrelate as a patient's health status changes. Typical queries include: which trajectory follows which other trajectory?, and how do trajectories evolve over a defined period of time? Such questions are particularly important in clinical practice and research, as they enable the detection of early warning signs, the mapping of disease progression patterns, and the evaluation of treatment effects. By modelling these temporal dependencies, this group provides a foundation for understanding the dynamics of health conditions rather than viewing them as isolated events.

The fourth set of questions, *AI/ML Techniques and Applications*, tests the ontology's capacity to represent computational models, data sources, and research artifacts. These questions evaluate whether the ontology can answer queries such as: which ML techniques are frequently used for disease prediction? What features are commonly employed in training such models? and which datasets and literature support these models? For instance, the ontology should link LSTM models with features like HbA1c levels and sleep patterns in the context of diabetes prediction. It should also allow users to identify relevant publications, such as studies proposing hybrid RNN frameworks for heart disease prediction, and associate these with datasets and input modalities (such as unstructured clinical notes, wearable device data).

5. Ontology Evaluation

To evaluate the practical utility of the ontology, we adopted a competency question–based approach, whereby the ontology was queried using SPARQL to test whether it could effectively answer real-world

analytical questions relevant to our domain. You can find evaluation done for all competency questions on the COPE website[1]. As clinical data is not readily available, we performed first level validation [35] using a synthetically generated healthcare dataset that simulates real-world EHRs including patients, encounters, conditions, medications, procedures, observations, care plans, payers, and claims [36].

In this section, we present the conceptual overview (in Figure 3), SPARQL query (in Listing 1) and the answers we obtained (in Table 2) for the competency question *Which patient characteristics are most strongly associated with increased susceptibility to specific diseases?*. Herewith, we aimed to validate whether the ontology could support the identification of key person-level characteristics associated with various chronic diseases and determine how frequently those characteristics appeared across individuals in the dataset instantiated in RDF. We also provide the conceptual overview of the two competency questions that focus on temporal aspects of COPE in Figures 4 and 5.

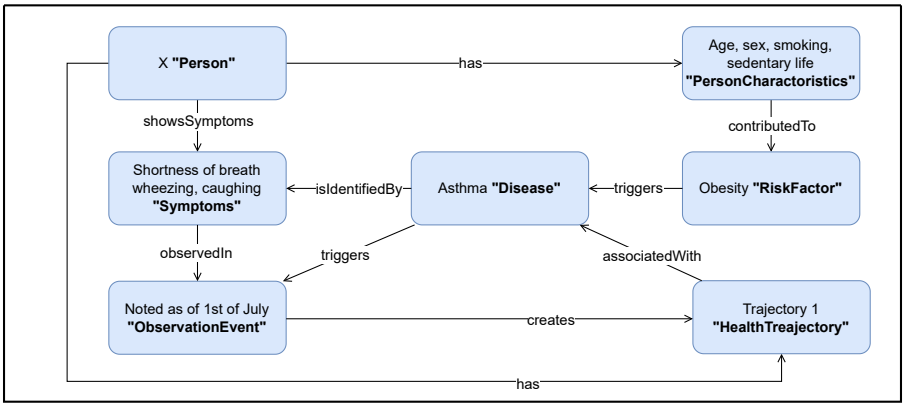

Figure 3: Conceptual overview of the competency question; *Which patient characteristics are most strongly associated with increased susceptibility to specific diseases?*

Table 2
A sample output showing person characteristics associated with diseases and the number of patients exhibiting them, based on the data available in the ontology implementation.

Person Characteristic	Disease Name	Patient Count
hasExerciseFrequency	Lung Cancer	4
alcoholConsumption	Lung Cancer	4
hasGender	Lung Cancer	4
hasAge	Lung Cancer	4
hasEthnicity	Lung Cancer	4
hasCountry	Lung Cancer	4
hasCity	Lung Cancer	4
hasHealthAttitude	Lung Cancer	4
hasGender	Hypertension	3
hasAge	Hypertension	3
hasHealthAttitude	Hypertension	3
hasEthnicity	Hypertension	3
hasExerciseFrequency	Hypertension	3
alcoholConsumption	Hypertension	3
hasCity	Hypertension	3
hasCountry	Hypertension	3
hasHealthAttitude	Diabetes Mellitus	2
hasHealthAttitude	Chronic Kidney Disease	2
hasPostalCode	Chronic Kidney Disease	1

[1]https://hzent.github.io/COPE/CopeDashboard.html#competency-questions

We executed a representative SPARQL query (shown in Listing 1) that retrieves all person-level characteristics linked to diseases via risk factors and aggregates the number of patients (such as RDF instances of `sio:Person`) exhibiting those characteristics. This query traverses the ontology's structure by combining several object properties—namely `COPE:hasCharacteristic`, `COPE:hasRiskFactor`, and `COPE:contributesTo`, and filters the results to include only relevant characteristic types (biological, demographic, behavioural, psychographic, and geographical) while excluding structural RDF elements such as `rdf:type`, `rdfs:label`, and `rdfs:comment`. The query aggregates patient counts grouped by both characteristic and disease, allowing us to gain insight into the most salient attributes influencing particular health outcomes.

```
PREFIX COPE: <https://purl.archive.org/cope#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT
  ?personCharacteristic
  ?diseaseName
  (STR(COUNT(DISTINCT ?person)) AS ?patient_count)
WHERE {
  ?person a sio:SIO_000498 ;
        COPE:hasCharacteristic ?charInd ;
        COPE:hasRiskFactor ?risk .

  ?charInd a ?charType ;
         ?personCharacteristic ?value .

  FILTER(?charType IN (
    COPE:Biological,
    COPE:Demographic,
    COPE:behavioural,
    COPE:Psychographic,
    COPE:Geographical))

  FILTER(
    ?personCharacteristic != rdf:type &&
    ?personCharacteristic != rdfs:label &&
    ?personCharacteristic != rdfs:comment)

  ?risk COPE:contributesTo ?disease .
  ?disease COPE:hasDiseaseName ?diseaseName .
}
GROUP BY ?personCharacteristic ?diseaseName
ORDER BY DESC(?patient_count)
```

Listing 1: SPARQL query to retrieve person characteristics associated with diseases and patient counts.

The results of this evaluation are shown in Table 2, which presents a subset of the query output. Notably, certain person characteristics such as `hasExerciseFrequency`, `alcoholConsumption`, `hasGender`, and `hasAge` emerged consistently across multiple disease types, particularly for Lung Cancer and Hypertension. For instance, Lung Cancer was associated with eight person-level features shared across four patient instances, indicating strong coverage of characteristic-disease relationships. Hypertension also showed high overlap, with similar attributes represented in three individuals. Furthermore, the ontology captured nuanced demographic and behavioural traits relevant to disease modelling. The properties `hasEthnicity`, `hasCity`, and `hasCountry` appeared frequently, highlighting the role of geo-social factors in health data representation. Psychographic traits such as `hasHealthAttitude` also featured across multiple diseases, validating the ontology's capacity to accommodate complex,

non-clinical risk factors. Importantly, the appearance of `hasPostalCode` in at least one patient confirms the ontology's support for fine-grained geospatial attributes.

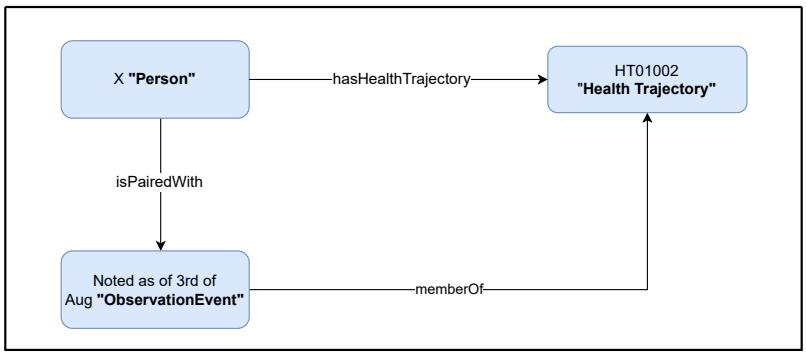

Figure 4: Conceptual overview of the competency question; *Which trajectory does the person participate in at a specific observation event?*

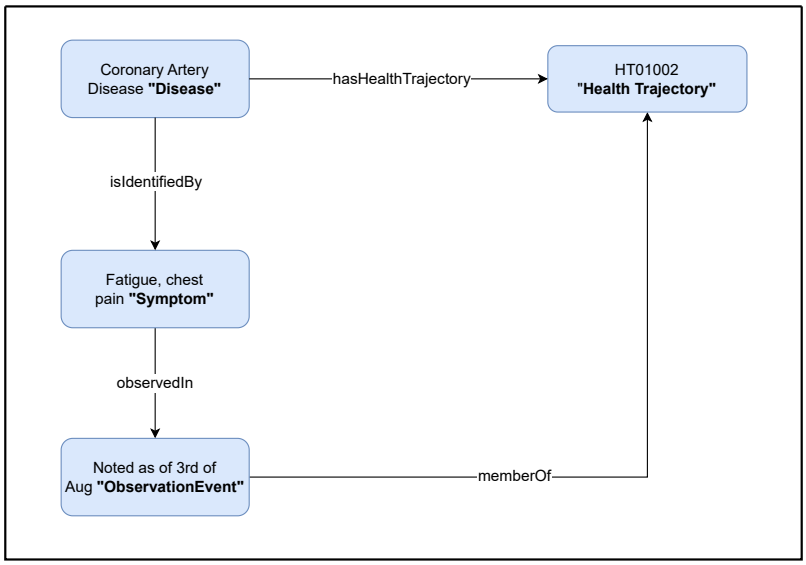

Figure 5: Conceptual overview of the competency question; *Which disease are the symptoms observed in within the current participating trajectory?*

In capturing temporal factors, Figure 4 illustrates the relationship between a person, their health trajectory, and observation events. A Person entity is linked to a Health Trajectory through the property `hasHealthTrajectory`, indicating the longitudinal record of their health status. Each person is also associated with an `ObservationEvent` via the property `isPairedWith`, which captures the temporal context of health-related data collection. The `ObservationEvent` is further connected to the corresponding Health Trajectory through the `memberOf` property, signifying that individual observations contribute to, and are embedded within, the broader health trajectory.

Similarly, Figure 5 depicts how a disease and its associated symptoms are linked within a health trajectory. The entity `Disease` (such as heart condition) is connected to a Health Trajectory via the property `hasHealthTrajectory`, indicating its progression over time. The disease is identified by symptoms such as fatigue and chest pain, which serve as clinical markers of its presence. These symptoms are further observed in a specific `ObservationEvent` , capturing the temporal context of their occurrence. The `ObservationEvent` itself is `memberOf` the Health Trajectory, thereby situating the recorded clinical findings within the patient's broader longitudinal health record. This representation

highlights the interplay between diseases, symptoms, and observation events, enabling structured tracking of clinical evidence across time.

COPE contains 36 classes and 48 relations, and we believe this is a sufficiently rich structure for modelling the target domain while remaining tractable for reasoning tasks. The reasoning complexity is non-trivial: as the ontology grows, entailment (especially with transitive and temporal relations such as `precedes`, `leadsTo`, `isPartOf`) and role chaining may increase the computational load. However, COPE has been structured to contain such complexity by limiting deeply nested axioms and by modularizing domain-specific expansions (for example, keeping clinical, device, and trajectory aspects relatively decoupled) to help mitigate blow-up. In terms of scalability, we anticipate that COPE should scale to integrate datasets containing tens of thousands to low hundreds of instances (patients, events, trajectories) without excessive reasoning latency, particularly for routine tasks like instance classification or SPARQL query answering. For very large-scale deployments (millions of individuals), one may adopt hybrid reasoning and query strategies (such as pre-computation, incremental reasoning, approximate reasoning, partitioning) to keep response times acceptable. Going forward, further empirical benchmarking (measuring reasoning time, memory usage, and query throughput) will be essential to validate these projections, especially when extending COPE with additional classes, axioms, or integrating with other ontologies.

6. Conclusion and Future Work

In this paper, we presented the first version of the Chronic Observation and Progression Events Ontology (COPE), an evolving, modular ontology designed to support the contextualized representation of person characteristics, observation events, and health trajectories. Grounded in domain expertise and literature, the ontology provides a semantic foundation for linking diverse concepts such as personal attributes, risk factors, symptoms, interventions, and outcomes in the context of disease progression.

This initial version serves as a basis for further refinement and extension. Our long-term vision is to leverage COPE in the development of digital twins for modelling individual health trajectories. Through such applications, the ontology can support personalized, AI-assisted decision-making in clinical and public health contexts. As our work progresses, we will iteratively enrich the ontology to reflect evolving data sources, computational models, and domain knowledge. Furthermore, while COPE has been designed with internally defined concepts such as Disease, Symptom, and Intervention to support first-level validation with synthetic clinical data, future work will focus on enhancing its interoperability with established biomedical standards. In particular, we plan to extend COPE by mapping its core concepts to widely adopted vocabularies such as ICD-10, SNOMED CT, and LOINC. This alignment will ensure semantic interoperability with EHR and existing biomedical ontologies, enabling COPE to integrate into clinical workflows and supporting consistent interpretation of healthcare information across systems. Also, we aim to explore how we can dynamically update the ontology with real-time model outputs or feedback loops from deployed systems.

Another avenue for future development is the incorporation of the Fast Healthcare Interoperability Resources (FHIR) framework. By aligning COPE with FHIR resources, we aim to support standardized data exchange and improve its compatibility with the growing ecosystem of FHIR-enabled healthcare applications. Such integration facilitates getting deeper insights into patients, diseases, observations, disease progressions, and their interrelationships, thereby enhancing the effectiveness of clinical decision support and operational healthcare applications. This extension will strengthen COPE's capacity to act as a bridge between patient data and clinical information systems, ensuring that the ontology not only serves research purposes but also supports real-world healthcare delivery and decision-making.

7. Declaration of Generative AI Use

Generative AI tools were not used in the creation of the intellectual or substantive content of this document. Grammarly was used for light editing, including grammar, spelling, and style suggestions.

References

[1] World Health Organization, Noncommunicable diseases, https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases, 2023. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

[2] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, New England Journal of Medicine 380 (2019) 1347–1358. doi:10.1056/NEJMra1814259.

[3] A. M. Alaa, M. van der Schaar, Forecasting individualized disease trajectories using interpretable deep learning, Nature Communications 10 (2019) 3923. doi:10.1038/s41467-019-11335-4.

[4] A. Senaratne, L. Seneviratne, Embedded to interpretive: A paradigm shift in knowledge discovery to represent dynamic knowledge, in: CEUR Workshop Proceedings, volume 3433, CEUR-WS, 2023.

[5] S. Chari, M. Qi, N. N. Agu, O. Seneviratne, J. P. McCusker, K. P. Bennett, A. K. Das, D. L. McGuinness, Making study populations visible through knowledge graphs, in: International semantic web conference, Springer, 2019, pp. 53–68.

[6] S. Chari, O. Seneviratne, D. M. Gruen, M. A. Foreman, A. K. Das, D. L. McGuinness, Explanation ontology: a model of explanations for user-centered ai, in: International semantic web conference, Springer, 2020, pp. 228–243.

[7] S. Chari, P. Acharya, D. M. Gruen, O. Zhang, E. K. Eyigoz, M. Ghalwash, O. Seneviratne, F. S. Saiz, P. Meyer, P. Chakraborty, et al., Informing clinical assessment by contextualizing post-hoc explanations of risk prediction models in type-2 diabetes, Artificial Intelligence in Medicine 137 (2023) 102498.

[8] S. Chari, O. Seneviratne, M. Ghalwash, S. Shirai, D. M. Gruen, P. Meyer, P. Chakraborty, D. L. McGuinness, Explanation ontology: A general-purpose, semantic representation for supporting user-centered explanations, Semantic Web 15 (2024) 959–989.

[9] N. Guarino, D. Oberle, S. Staab, What is an ontology?, Handbook on ontologies (2009) 1–17.

[10] N. F. Noy, D. L. McGuinness, Ontology Development 101: A Guide to Creating Your First Ontology, Technical Report KSL-01-05, Stanford Knowledge Systems Laboratory, 2001. https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.

[11] R. Cornet, N. de Keizer, Forty years of snomed: a literature review, BMC Medical Informatics and Decision Making 8 (2008) S2. doi:10.1186/1472-6947-8-S1-S2.

[12] C. J. McDonald, S. M. Huff, J. G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook, et al., Loinc, a universal standard for identifying laboratory observations: a 5-year update, Clinical chemistry 49 (2003) 624–633.

[13] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, J. A. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al., Gene ontology: tool for the unification of biology, Nature genetics 25 (2000) 25–29.

[14] L. M. Schriml, C. Arze, S. Nadendla, Y.-C. Chang, M. Mazaitis, V. Felix, G. Feng, W. A. Kibbe, C. Greene, Human disease ontology 2018 update: classification, content and workflow expansion, Nucleic acids research 47 (2019) D955–D962.

[15] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, S. Mundlos, The human phenotype ontology: a tool for annotating and analyzing human hereditary disease, American journal of human genetics 83 (2008) 610–615.

[16] J. Hobbs, F. Pan, Time ontology in owl, W3C Working Draft, 2006. https://www.w3.org/TR/owl-time/.

[17] F. Li, J. Du, Y. He, H.-Y. Song, M. Madkour, G. Rao, Y. Xiang, Y. Luo, H. W. Chen, S. Liu, et al., Time event ontology (teo): to support semantic representation and reasoning of complex temporal relations of clinical events, Journal of the American Medical Informatics Association 27 (2020) 1046–1056.

[18] R. Chandra, S. Agarwal, S. S. Kumar, N. Singh, Ocep: An ontology-based complex event processing framework for healthcare decision support in big data analytics, arXiv preprint arXiv:2503.21453 (2025).

[19] O. Seneviratne, A. K. Das, S. Chari, N. N. Agu, S. M. Rashid, J. McCusker, J. S. Franklin, M. Qi, K. P.

Bennett, C.-H. Chen, et al., Semantically enabling clinical decision support recommendations, Journal of Biomedical Semantics 14 (2023) 8.
[20] G. C. Publio, A. Ławrynowicz, L. Soldatova, P. Panov, D. Esteves, J. Vanschoren, T. Soru, Ml-schema: An interchangeable format for description of machine learning experiments, Semantic Web 0.0 (2020) (2020) 1–11.
[21] P. Panov, L. Soldatova, S. Džeroski, Ontology of core data mining entities, Data Mining and Knowledge Discovery 28 (2014) 1222–1265.
[22] L. Zhang, W. Zhao, W. Gao, W. Yu, Z. Song, Ontology-driven decision support systems in healthcare: a review, Artificial Intelligence in Medicine 124 (2022) 102195.
[23] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.
[24] M. Dabrowski, M. Synak, S. R. Kruk, Bibliographic ontology, in: Semantic digital libraries, Springer, 2009, pp. 103–122.
[25] S. Peroni, D. Shotton, The spar ontologies, in: The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17, Springer, 2018, pp. 119–136.
[26] H. M. Krumholz, P. M. Currie, B. Riegel, C. O. Phillips, E. D. Peterson, R. Smith, C. W. Yancy, D. P. Faxon, A taxonomy for disease management: a scientific statement from the american heart association disease management taxonomy writing group, Circulation 114 (2006) 1432–1445.
[27] N. Halfon, M. Hochstein, Life course health development: an integrated framework for developing health, policy, and research, The Milbank Quarterly 80 (2002) 433–479.
[28] G. Bediang, G. Falquet, A. Geissbuhler, An ontology-based semantic model for sharing and reusability of clinical pathways across context (share-cp), in: MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation, IOS Press, 2022, pp. 86–90.
[29] A. Senaratne, L. Seneviratne, Anomaly detection for prolongation of health, https://extravaganza.gallery/dhw25/exhibits/88ed8446-8931-44e7-b702-c45416995f37, 2025. Accessed: 24 July 2025.
[30] L. Seneviratne, A. Senaratne, Data atomization: A framework for on-demand association and access control of sensitive data, in: 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), IEEE, 2025, pp. 2275–2280.
[31] A. Senaratne, Anomaly detection in graphs for knowledge discovery and data quality enhancement (2024).
[32] A. Senaratne, P. Christen, G. Williams, P. G. Omran, Rule-based knowledge discovery via anomaly detection in tabular data, in: CEUR Workshop Proceedings, volume 3433, CEUR-WS, 2023.
[33] A. Senaratne, P. Christen, G. Williams, P. G. Omran, Unsupervised identification of abnormal nodes and edges in graphs, ACM Journal of Data and Information Quality 15 (2022) 1–37.
[34] D. Garijo, Widoco: a wizard for documenting ontologies, in: International Semantic Web Conference, Springer, Cham, 2017, pp. 94–102. URL: http://dgarijo.com/papers/widoco-iswc2017.pdf. doi:10.1007/978-3-319-68204-4_9.
[35] J. Chen, D. Chun, M. Patel, E. Chiang, J. James, The validity of synthetic clinical data: a validation study of a leading synthetic data generator (synthea) using clinical quality measures, BMC medical informatics and decision making 19 (2019) 44.
[36] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, S. McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, Journal of the American Medical Informatics Association 25 (2018) 230–238.