Group 16 Masters Spring Research Project 2023-2024: Automated Annotation of Tabular Data for Clinical Applications

Anton Changalidi, Aurora Pia Ghiardelli, Ashkan Karimi Saber, Ivan Poliakov

June 2024

Contents

Abstract

This research presents the development of an annotated dataset derived from the MIMIC-III Demo dataset in tabular form. Comprehensive annotations at multiple levels, including column types, cell entities, and column properties, were systematically linked to standard medical ontologies. These detailed annotations facilitate the precise identification and classification of tabular elements within the medical domain. The annotated tables could serve as a valuable resource for training and evaluating automated annotation tools. The aim of this work is to establish a ground truth annotated dataset that could be effectively utilized in conjunction with the aforementioned tools, enhancing the interoperability of medical records in healthcare and research settings.

1 Introduction and Related Work

A table is the most common and organized way to represent data, making the tasks of accessing, manipulating, and extracting information straightforward due to its structured nature. However, reading a table without prior understanding can be challenging. Annotated tables, enhanced with details on the semantic significance of their columns and their interconnections, and mapped onto a standardized ontological reference, address this issue. Despite column labels offering some insight, their linkage to standard medical ontologies might not be immediately apparent. This is especially true when working with data that follows an unfamiliar administrative standard or is written in a foreign language.

In the era of Web 3.0 [1], the volume and diversity of tabular data underscore the necessity for methods that can automatically provide such annotations about the tables. This is why the field of Semantic Table Annotation has gained so much attention in recent years [1]. Besides generally simplifying the workflow with tabular data, annotated tables significantly improve readability and prove invaluable for merging tables in different formats or languages. It becomes even more important in a clinical context, where data can be challenging to read and interpret without considerable prior knowledge.

Additionally, in situations where relational data holds significant value, the ability to extract and convert the relationships between entities within a table into relational data can offer substantial benefits. Temporal, quantitative, and unit-based features are more effectively interpreted when contextualized with the specific attributes they characterize.

In automatic table annotation, there exists a diverse range of methodologies. Among them, some methods fall into the category of symbolic approaches, which are based on the identification and interpretation of syntactic and semantic structures within tables using predefined rules or probabilistic language models. More recent approaches take advantage of machine learning embedding power to generate vector representations of table instances and relations, effectively capturing their semantic meaning [2]. Solutions involving a combination of the inferential power of symbolic reasoning with the predictive capacity of neural networks are referred to as neuro-symbolic AI [2]. Due to the complementary nature of tabular and ontological data - while tabular data is specific and provides sparse entities with few relations, knowledge graphs provide non-specific but comprehensive models of entities and relations. In this research we particularly focused on automated table annotation method called MUT2KG [2]. This is an algorithm based on a well-known energy-based embedding translation model TransE. MUT2KG embeds the entities and relations derived from supporting knowledge graphs in alignment with tabular data, enriching the embedding space while maintaining the relational integrity of table elements and graph nodes within a vector space.

However, while applying the MUT2KG algorithm in less specific domains is feasible using knowledge graphs such as DBpedia[3] or Wikidata[4], its application in clinical domains requires prior creation of a standardized knowledge graph with a verified schema representative of hospital protocols. Accomplishing this is not a trivial task, as it necessitates meticulous manual work and relies on international vocabularies and correct schema.

This research will contribute to the generation of a representative clinical ontology of EHRs that will be useful both for implementing the MUT2KG algorithm in clinical data and for constructing a reference knowledge graph for the healthcare system. It will analyze how knowledge graphs can be beneficial in the healthcare domain, their potential, and the reasons why they have not been widely implemented in many contexts, despite their advantages. The pipeline of the MUT2KG algorithm will be explained for the main tasks of tabular annotation (types, properties, and entities annotations, or CTA, CPA, and CEA respectively). Furthermore, the main contribution of the research project will be explained, which involves generating an ontology based on the MIMIC III dataset. This process will involve manual annotation of both table elements and relations among these elements, using reference knowledge bases such as SNOMED-CT(Systematised Nomenclature of Medicine Clinical Terms), LOINC(Logical Observation Identifiers Names and Codes), ICD-10(International Classification of Diseases 10th Revision) and other international terminologies. This effort will not only aid in creating a knowledge graph usable in the neuro-symbolic pipeline of the MUT2KG algorithm but also associate labels with table elements for supervised learning and enhance embedding learning models. Once the manual annotation process is completed, these annotations can be used to create a knowledge graph representing the ontology of the pipeline. This knowledge graph can be generated using declarative graph generation techniques such as SDM-RDFizer[5].

1.1 Context

The main goal of the research presented in this article is to create an annotated tabular dataset in a clinical context. This could further allow for the implementation of the MUT2KG[2] algorithm pipeline with clinical data. Before explaining the actual contribution of this research, it is important to analyze what it means to deal with clinical data and why applying this pipeline in that domain is not straightforward.

Clinical data are encapsulated within Electronic Health Records (EHRs), which serve as the digital counterparts to traditional patient hospital records. EHRs are exceptionally informative, encompassing a comprehensive array of data including patient demographics, family medical history, diagnostic information, procedural details, medication administration records, and administrative data pertaining to hospital locations visited and healthcare professionals involved in patient care [6].

The fact that they cover all the information about the patients during their hospital stay makes them highly informative for researchers; indeed, they can be used for epidemiological and health studies, identifying risk factors, tracking disease progression, and analyzing therapy performance [6]. Additionally, they are valuable for doctors who are on the front line and need to make decisions regarding the patient;

The main downside of this type of data is that they lack a consistent structure, vary depending on their source, and are often disorganized and difficult to analyze [6]. Additionally, when organized in a tabular format, they lack an obvious relational structure, which complicates the consideration of all important relations in the decision-making process. Furthermore, these data often include numerous numerical values and codes that outsiders may find difficult to identify, necessitating a manual annotation process that incurs significant costs.

All these examples demonstrate the potential benefits of applying automatic annotation to

Electronic Health Records (EHRs) to obtain a semantic interpretation of the tables.

As mentioned previously, in MUT2KG the elements of the table are aligned with the elements of a reference ontology represented in the form of a knowledge graph. For general domains, there are ontologies such as DBPedia[7] or Wikidat a[4], however, the fundamental issue in the process of approaching a clinical context is that there is currently no comprehensive reference knowledge graph containing terms of medical knowledge, drugs, interactions between them, clinical protocols, and real clinical data. Therefore, the creation of this reference knowledge graph would not only help hospitals in data organization and the decision-making process, but it would also enable the application of the neuro-symbolic automatic annotation pipeline in the clinical context.

1.2 Knowledge Graphs in Health Care Domain

As previously mentioned, the EHRs system is the main format in which all patient information is recorded and transcribing this data into a relational structure (e.g. the knowledge graph) would offer numerous benefits. One of the main advantages of knowledge graphs is their ability to integrate current data with information from other sources, thereby enhancing the informativeness of the data structure[6]. This relational structure will allow for organized integration and easy interpretation for the observer, enabling simplified navigation through these data sources.

Another advantage of knowledge graphs lies in their explicit visualization of relationships between nodes. Having a structure that explicitly expresses relationships among elements allows for simplified navigation across nodes, making even unknown information accessible to the observer. This is particularly crucial in diagnostic processes where knowledge is causal [8], where having a data structure that facilitates the visualization of interactions between elements might be advantageous. Additionally, having a reference knowledge graph would help researchers and doctors relate the information in a patient's medical record to that of other patients with similar clinical characteristics, narrowing down the domain of potential diagnoses and treatments. Furthermore, having a standard organizational schema, represented by a knowledge graph, would assist hospitals in organizing clinical data by providing a standardized framework.

One example of the potential of knowledge graphs in the healthcare domain is presented in the article [9], where knowledge graphs are used to determine new gene-disease associations. Another example is shown in the studies [10, 11] where knowledge graphs were used in the prediction of heart failure patients; whether in another research [12] graphs were used in the monitoring of COVID-19 spread.

Despite the positive changes that using a graph data structure to organize hospital records would bring, there are few knowledge graph models for clinical data, and those that exist do not interconnect different data sources and domains but tend to be domain-specific [6]. The cause of this lies not only in the difficulty of transforming a large quantity of tabular data into RDF(Resource Description Framework) format but also in establishing the correct relational structure underlying this data. In fact, there is no rigorous evaluation model for these knowledge graph models [6], making the transition into such a delicate domain as healthcare challenging.

One of the notable attempts to create a relational structure for hospital records is represented by the Swiss Personalized Health Network(SPHN) initiative [13]. The idea behind this initiative is to establish a reference ontology for hospital clinical data, aiming to address the interoperability challenges between healthcare, research, and regulatory agencies [14]. By defining a relational schema of building block concepts for organizing healthcare data linked to a set of international vocabularies such as SNOMED International, LOINC, or ICD-10, the understanding and analysis of clinical data are facilitated.

SPHN seeks to create a standardized framework that enables seamless integration and exchange of clinical information across different stakeholders in the healthcare ecosystem. This approach not only enhances data interoperability but also supports advanced analytics and research efforts by providing a structured foundation for understanding and utilizing healthcare data effectively.

1.3 TransE and MUT2KG

There exist a number of strategies for the problem of automated tabular data annotation. The approach highlighted in this research project is MUT2KG. It is a neuro-symbolic pipeline that uses TransE[15], a well-known algorithm in the field of knowledge graph embedding, to learn vector representations of cells/nodes and relations between them for both tabular and knowledge graph, with subsequent alignment of both vector spaces.

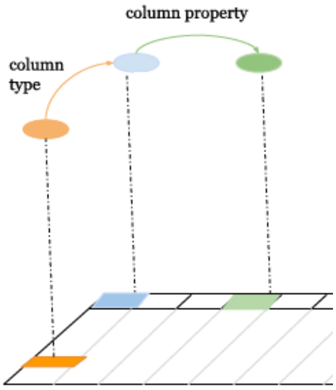


Figure 1: CTA, CEA, and CPA tasks in a table [2]

Automatic table annotation involves a multi-class classification task (Figure 1), aiming to classify column types, properties between columns and between table cells, and table cells. CTA (Column Type Annotation) is used to specify the type or class of the columns in a table. CPA (Column Property Annotation) provides additional information about the relationships of the columns. CEA (Cell Entities Annotation) involves annotating individual cells of a table, specifying the entities these cells represent. The MUT2KG algorithm[2] conducts this classification task using the TransE knowledge graph embedding algorithm. The pipeline of this algorithm is shown in the Figure 2.

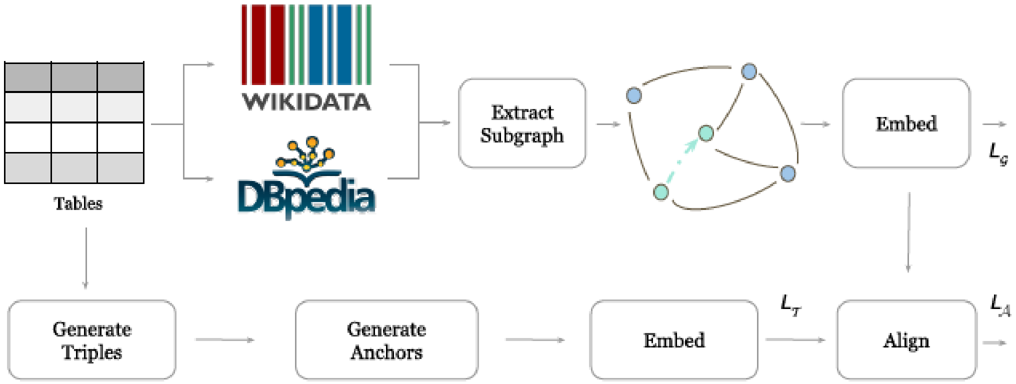


Figure 2: MUT2KG pipeline [2].

The algorithm uses Knowledge Graph Embedding techniques that generate vector representations of graph elements based on the idea that the vector of an object $\vec{t}$ in a relation should be equal to the vector of the subject $\vec{h}$ plus the vector of the relation $\vec{p}$. By using this constraint $\vec{h} + \vec{p} = \vec{t}$ in the determination of a loss function for an entire knowledge graph, it is possible to obtain vector representations of nodes and relationships that seek to minimize this function. Using this schema TransE algorithm[15] generates vector representations based on the relational structure of individual nodes rather with no natural language learning, while allowing close representations even for names in different formats or languages.

In relation to the table to be annotated, the same approach is used, extracting triplets (e_1, r, e_2) from it such that each column is related to another column with unique relationships, and each cell is related to every cell in the same row with the same relationships that connect the columns. Once these triplets are obtained, the TransE algorithm can be applied to them, resulting in vector representations of the table columns, cell elements, and relationships between columns and cells of the table.

The TransE algorithm uses the following loss function in learning the embeddings of the elements of the table:

$$L_{\mathcal{T}} = \sum_{(e_1,r,e_2)\in\mathcal{T}} \|\vec{e_1} + \vec{r} - \vec{e_2}\|^2$$

and the following loss function in learning the elements of the graph:

$$L_{\mathcal{G}} = \sum_{(h,p,t)\in\mathcal{G}} \|\vec{h} + \vec{p} - \vec{t}\|^2$$

Then, the loss function for aligning the elements of the table $\mathcal{T}$ with elements of the graph $\mathcal{G}$ is represented by the formula:

$$\mathcal{L}_{\mathcal{A}} = \sum_{(a,e)\in\mathcal{A}} \sum_{(a,e')\notin\mathcal{A}} \left(\|\vec{a} - \vec{e}\|_2^2 - \|\vec{a} - \vec{e'}\|_2^2\right) + \gamma$$

where $(a, e) \in \mathcal{G} \times T$ is an anchor pair that relates column elements with graph ontology concepts and γ is a radius hyperparameter around the anchor nodes. The total loss function of the pipeline is computed as the sum of these loss functions:

$$L_{\text{total}} = L_{\mathcal{G}} + L'_{\mathcal{T}} + L'_{\mathcal{A}}$$

In addition, it is possible to use Natural Language Processing techniques such as BERT[16] or DistilBERT[17] to incorporate a linguistic component into the vector representations.

The experiments regarding this pipeline were conducted on the tabular data proposed by the SemTab 2023 Challenge and reported very good precision results on Wikidata tables for the property classification task (0.948), while yielding less satisfactory results for entity prediction (CEA) (0.587) and type prediction (CTA) (0.655). Experiments were also conducted regarding MUT2KG-II, a superior sibling of MUT2KG-I that incorporates improvements in terms of loss function; experiments with this latter method led to performance enhancements, achieving higher precision levels across all datasets and tasks [2].

1.4 Contribution

The current research aims to contribute to the creation of a representative knowledge graph of EHRs, enhancing the MUT2KG algorithm for semantic annotation of hospital data. This will be achieved by generating a knowledge graph from the benchmark dataset MIMIC III, a publicly available relational dataset comprising over forty thousand patients who stayed in critical care units at Beth Israel Deaconess Medical Center.

To generate a knowledge graph from this benchmark dataset, it is necessary to translate the dataset into standardized and international vocabularies (such as SNOMED International, LOINC, or ICD-10) and use reference schemas, like SPHN to extract both the main ontologies of the reference dataset and the relations between them. This will be accomplished through a manual annotation process of the dataset, from which a knowledge graph can then be generated. This knowledge graph can not only be used in the tabular annotation pipeline but can also serve as a representative clinical ontology in RDF format for clinical data. Furthermore, as mentioned in the section 1.2, a major advantage of knowledge graphs is their possibility of being expanded using other data sources; this will thereby enhance the information capacity and domain coverage of the current ontology.

The annotation procedure will follow the same tasks of automated tabular annotation, namely CTA, CEA, and CPA. Subsequently, this manual annotation process will also serve as a benchmark for automatic annotation, facilitating supervised learning concerning the MIMIC dataset and aiming to enhance the MUT2KG algorithm.

An additional contribution of this research will involve the analysis and navigation of the ontologies and knowledge bases used in the manual annotation of the relational tables within the MIMIC dataset. This manual annotation procedure will allow for evaluating how these ontological schemas fit clinical hospital data. Specifically, it will assess how the relationships specified in the SPHN RDF Schema align with the tabular structure of the MIMIC dataset, analyzing whether the schema can annotate a significant portion of the relationships between columns.

1.5 Research Questions

Integrating clinical ontologies with Electronic Health Records (EHR) data is essential for advancing healthcare analytics and decision support systems. As the volume of health data grows, effective semantic annotation becomes crucial for improving data interpretability and utility. The following questions examine the integration of clinical ontologies into EHR systems, assessing their benefits, practicality, representativeness, and verification processes:

- **Utility of Clinical Ontology Integration: Why can a clinical ontology integrated with EHR data be useful?** This question examines the enhancements in healthcare quality and cost reduction through better risk identification, disease tracking, and treatment effectiveness assessment using annotated EHRs.

- **Representativeness of Clinical Ontologies: What are the main features of the currently available clinical ontologies? Are they sufficiently representative of hospital data?** This inquiry evaluates whether these ontologies, such as SNOMED and LOINC, are comprehensive enough to cover the data typically found in hospital records, ensuring they can support various clinical and administrative needs.

- **Practicality of Ontology Use in Annotations: Is it possible to effectively use these ontologies for manually annotating a clinical table?** This question focuses on the challenges and possibilities of applying ontological frameworks to the structured annotation of clinical data tables, considering the specific context of healthcare environments.

- **Verification of Annotation Correctness: Once this annotation is done, how can we ensure its correctness?** This question addresses the methods and strategies needed to validate the accuracy and reliability of the annotations, ensuring they meet the required standards for clinical use and research.

By addressing these questions, this research aimed to not only enhance the theoretical understanding of semantic annotation in healthcare but also to provide practical insights that can lead to more effective and reliable use of EHR data in clinical environments. The outcomes of this investigation are expected to contribute significantly to the field of health informatics, supporting advancements in data-driven healthcare solutions.

2 Data

The data collection phase of the project involved meticulous selection and annotation of electronic health records (EHR) data to ensure a representative and useful dataset for the semantic annotation research. A single patient's data from the MIMIC III database was selected at random (identified by $subject\_id = 42412$) to ensure a balanced representation across various medical tables. This approach allowed us to maintain consistency in data volume and complexity across the tables that contained the $subject\_id$ and those that did not, ensuring a comprehensive dataset.

The annotation process was methodically structured to improve readability and utility. A combination of CTA, CEA, CPA was used. For instance, annotations were integrated directly alongside the MIMIC tabular data. First CTA and CEA were defined for each column, followed by CPA links and other relational triples that involved a predicate, laid out below the initial tabular data. This structure facilitated a clearer understanding of the relationships within the data, essential for the neuro-symbolic semantic analysis. This is shown in Figure 1.

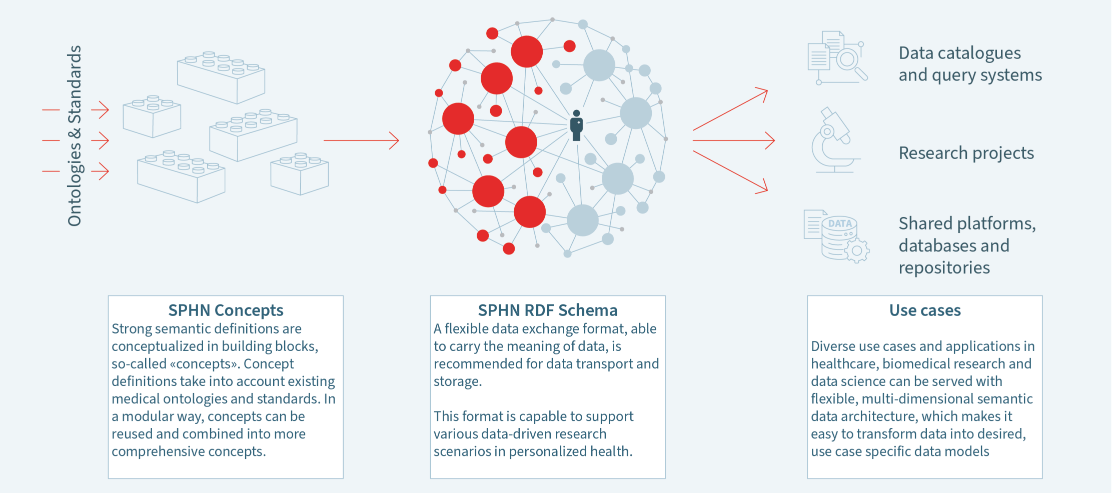
Moreover, the data annotation strategy involved addressing several challenges and questions related to the annotation process, such as the appropriateness of using row_id for linking data and the decision-making around annotating columns with missing values or those with less obvious semantic connections. These considerations were crucial for ensuring the accuracy and relevance of annotations, which are pivotal for the subsequent analysis phases of the current project.

In summary, the data collection and annotation phase was foundational for the current research, setting the stage for analyzing the effectiveness of clinical ontologies in enhancing the interpretability and utility of EHR data. Through this rigorous process, a robust dataset was established that would support the investigations into the integration of neuro-symbolic AI with clinical data analytics, driving forward the capabilities in healthcare informatics.

2.1 Knowledge Basis and Schema

When generating a reference ontology in RDF format from a benchmark dataset, it is important to link table elements to standard terminologies that enable unique identification within this ontology. Furthermore, in constructing a relational structure, it is necessary to reference verified schema that accurately reproduces hospital protocols.

As the primary schema, the SPHN RDF Schema, created in 2017[13] to enhance interoperability between healthcare and research (the strategy of the SPHN initiative is shown in Figure 3), has been chosen. This schema provides a relational framework for a range of fundamental healthcare concepts linked to international vocabularies representing these domains(external terminologies). Queryable through an interactive tool, it allows querying both types and relationships among these types. However, the graph does not contain data on real patients, which would be beneficial for cell entity annotation. Furthermore, although this schema represents a reliable relational structure for column property annotation purposes, it does not cover all possible relationships between columns. Additionally, many columns in the MIMIC dataset did not have corresponding entities in the SPHN schema, complicating the annotation process.


Ontologies & Standards
Data catalogues and query systems
Research projects
DATA
Shared platforms, databases and repositories

SPHN Concepts
Strong semantic definitions are conceptualized in building blocks, so-called «concepts». Concept definitions take into account existing medical ontologies and standards. In a modular way, concepts can be reused and combined into more comprehensive concepts.

SPHN RDF Schema
A flexible data exchange format, able to carry the meaning of data, is recommended for data transport and storage.

This format is capable to support various data-driven research scenarios in personalized health.

Use cases
Diverse use cases and applications in healthcare, biomedical research and data science can be served with flexible, multi-dimensional semantic data architecture, which makes it easy to transform data into desired, use case specific data models

Figure 3: SPHN semantic strategy [13]

In the following sections, it will be shown how the schema was queried to proceed with the annotation of the types and properties of the benchmark dataset.

For the task of entity annotation, internationally recognized vocabularies such as ICD-10 for diagnoses and the Anatomical Therapeutic Classification (ATC) were used. Additionally, the international SNOMED-CT browser, recognized and used in 50 countries worldwide, was employed. In the following sections, the procedure for the task of cell entity annotation will be explained.

3 Methods and Results

In the current research, the semantic annotation of tables, a process that transforms raw data into a more interpretable format, consists of three main tasks: CTA, CPA, and CEA. These tasks collectively ensure that each column, cell, and their relationships within a table are precisely understood and contextualized. In the following subsections, further details about these tasks are given.

3.1 CTA

Column Type Annotation (CTA) refers to the process of annotating each column of a table with the most appropriate type or concept from a predefined ontology. This helps in understanding the nature of data contained within the column and its relation with other data points. the reasons for the importance of this task are:

1. Interoperability: By standardizing the semantic types of columns, data from different sources can be integrated and used together more effectively.
2. Data Quality and Consistency: Ensures that the data adheres to a known structure, improving data quality and consistency.
3. Automated Reasoning: Facilitates automated reasoning and inference over the data by providing clear semantic meanings.

3.1.1 Methodology

The main goal of the CTA task is to annotate each column in a table with a type from a predefined ontology, and in order to accomplish this, these steps were carried out:

- **Understanding the data:** In this step, the contents of the table were examined to understand what kind of data is present in each column. This involves looking at the column headers, sample values, and the overall context of the table.
- **Assigning Semantic Types:** After Choosing an appropriate ontology that covers the domain of the data (in this research, SPHN and SNOMED), for each column, the concept in the ontology that best describes the data in that column was determined.
- **Validation:** To systematically validate the type annotations of the columns, tables were assigned to two annotators for independent execution of the Column Type Annotation (CTA) task. Subsequently, their annotations were compared to ascertain consistency. In instances of discrepancies, a third annotator was engaged to review the conflicting labels and determine the definitive annotation. This methodology significantly enhanced the accuracy and reliability of the CTA task outcomes.

3.1.2 Examples

In this section, some examples for the CTA task are given below: In table 1 part of the table ADMISSION[1] from MIMIC dataset with 3 columns and their corresponding CTAs is seen.

admission_location	**discharge_location**	**insurance**
Location	Location	HealthInsurance

Table 1: Admission table and its CTA.

[1] `https://mimic.mit.edu/docs/iii/tables/admissions/`

3.2 CPA

Column Predicate Annotation (CPA) is the process of assigning appropriate properties to the relations between pairs of columns in a table, the methodology for this is explained in the following section.

3.2.1 Methodology

The methodology for CPA involves establishing relations between column types within the dataset. This process was guided by the structured framework of the MIMIC III dataset, wherein CPA aims to capture the predicates that link properties of one column to another, facilitating a relational understanding of the data. For instance, in the annotations, relations such as *item_id+hasStartDateTime* → *starttime* and *icustay_id+hasAdministrativeCase* → *hadm_id* were identified and mapped. These mappings are crucial for representing the semantics of the data, allowing for more complex queries and analyses that reflect the clinical context more accurately. The current approach also involved cross-referencing with clinical ontologies to ensure that the relationships are not only syntactically correct but also semantically meaningful.

3.2.2 Examples

Direct relations are the vast majority of the CPA. The examples are:

- **Input Events** table: *hadm_id* (AdministrativeCase) to *icustay_id* (CareHandling) with property hasCarehandling;
- **Prescriptions** table: *drug* (Drug) to *drug_type* with property hasCategory, to *drug_name_poe* and *drug_name_generic* with property hasName (it has 2 types of names names).

The more complex type of relation is an indirect one. The following list provides some examples of such CPAs as implemented in the dataset:

- **OutputEvents Table:** *rowid* → *itemid*, was implemented as *rowid* (Assessment Event) to Assessment with hasAssessment property and then to *itemid* (Assessment Result) with hasResult property (Figure4).
- **OutputEvents Table:** *itemid* → *value&valueuom*, was implemented as *itemid* (Assessment Result) to Quantity with hasQuantity property and then to *value* (Double) and *valueuom* (Unit) with hasValue and hasUnit respectively demonstrating the relation between medical measurements and their respective units. This helps in understanding the quantity and the metric used directly within clinical assessments.

These examples illustrate the practical application of CPA within the dataset, showcasing how property annotations enhance the interpretability and usability of EHR data for clinical research and practice. Each example serves as a demonstration of how complex relationships between healthcare data elements are semantically structured to support enhanced data-driven decision-making in healthcare settings.

3.3 CEA

3.3.1 Methodology

Cell Entity Annotation (CEA) enhances the specificity of data in the EHR dataset, addressing challenges where column types alone provide insufficient information, such as in the case of specific drugs. To achieve precise annotations, the following were utilized:

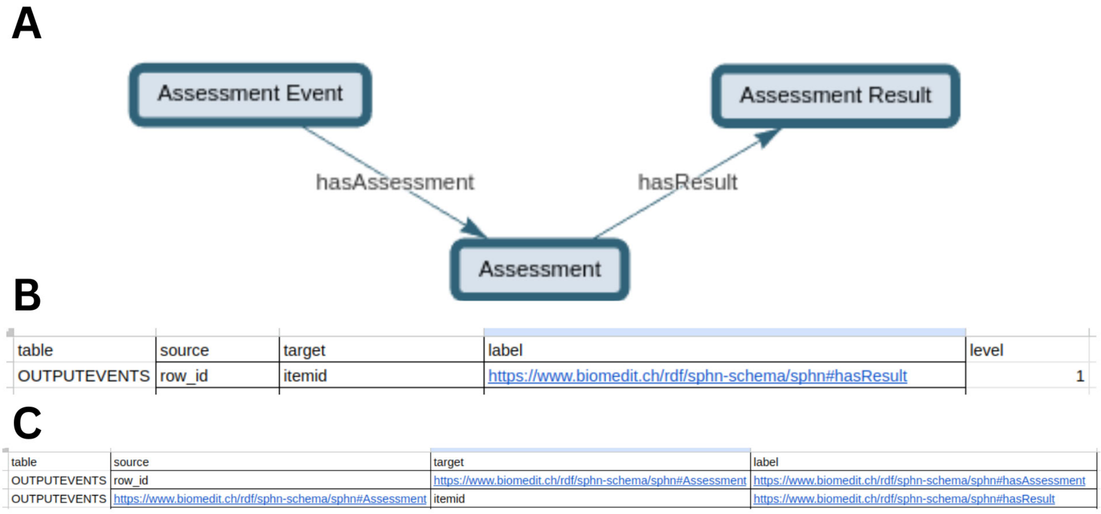


table	source	target	label	level
OUTPUTEVENTS	row_id	itemid	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasResult	1

table	source	target	label
OUTPUTEVENTS	row_id	https://www.biomedit.ch/rdf/sphn-schema/sphn#Assessment	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasAssessment
OUTPUTEVENTS	https://www.biomedit.ch/rdf/sphn-schema/sphn#Assessment	itemid	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasResult

Figure 4: CPA representation for Output Events Event to Assessment Result through intermediate node: (A) graph representation; (B) short format; (C) Long format.

- SNOMED CT (clinical terms) [18]: This comprehensive healthcare terminology system provides standardized medical terms for various healthcare aspects, improving data interoperability and clinical reporting.

- PubChem [19]: This database was used for detailed chemical and pharmaceutical data. It was essential for annotating specific drug entities, linking them accurately to their chemical identities.

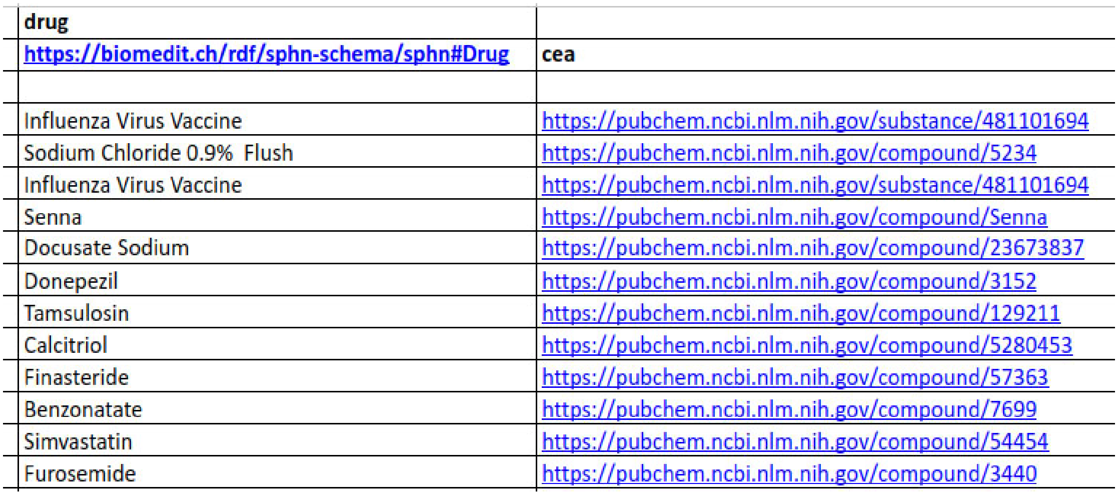
By integrating SNOMED CT and PubChem, CEA methodology of the current work effectively bridges the gap between general column types and the need for specific medical annotations, enhancing data analysis and supporting clinical decision-making.

3.3.2 Examples

The following list provides examples from the dataset, demonstrating the application of CEA with specific references to PubChem and SNOMED, using links for direct reference:

- **Prescriptions Table:** For the *drugs*, rather than just listing it under a general drug category, it was linked to its specific chemical composition available on PubChem. This allows for a more detailed understanding of its properties and usage. Examples are listed in the Figure 5

- **Input Events Table, status descriptions:** there are examples of possible status and its annotation:
  - **Rewritten:** Linked to SNOMED Changed status .
  - **Stopped:** Annotated with the SNOMED Stopped work.
  - **Finished Running:** Linked to SNOMED successfull status.



drug	
https://biomedit.ch/rdf/sphn-schema/sphn#Drug	cea
Influenza Virus Vaccine	https://pubchem.ncbi.nlm.nih.gov/substance/481101694
Sodium Chloride 0.9% Flush	https://pubchem.ncbi.nlm.nih.gov/compound/5234
Influenza Virus Vaccine	https://pubchem.ncbi.nlm.nih.gov/substance/481101694
Senna	https://pubchem.ncbi.nlm.nih.gov/compound/Senna
Docusate Sodium	https://pubchem.ncbi.nlm.nih.gov/compound/23673837
Donepezil	https://pubchem.ncbi.nlm.nih.gov/compound/3152
Tamsulosin	https://pubchem.ncbi.nlm.nih.gov/compound/129211
Calcitriol	https://pubchem.ncbi.nlm.nih.gov/compound/5280453
Finasteride	https://pubchem.ncbi.nlm.nih.gov/compound/57363
Benzonatate	https://pubchem.ncbi.nlm.nih.gov/compound/7699
Simvastatin	https://pubchem.ncbi.nlm.nih.gov/compound/54454
Furosemide	https://pubchem.ncbi.nlm.nih.gov/compound/3440

Figure 5: CEA representation for Prescription's Drug.

The approach, that was used in this section, efficiently managed large datasets containing thousands of rows by using a Python dictionary for automated annotation. This method assigned unique keys to specific terms, linking them to their respective SNOMED or PubChem identifiers, greatly enhancing the speed and accuracy of the process. This streamlined method demonstrates a scalable and precise solution for extensive EHR data, essential for maintaining consistency and reliability in healthcare informatics.

4 Discussion

During this research, annotations for a subset of the MIMIC III Demo dataset were developed, focusing on column type, column property, and cell entity annotations. The work was restricted to records of a single patient, chosen due to the substantial number of records available, which facilitated meaningful entity annotations. The annotations for column type and column property are applicable universally across all patient records in the MIMIC dataset. The task of cell entity annotation remains, which can be semi-automated as demonstrated during the analysis of the selected patient's data. It was observed that some tables could contain over a thousand rows related to just this single patient, necessitating partial automation of the cell entity annotation process as detailed in the methodology section.

The approach to cell entity annotation has limitations, particularly for features requiring manual intervention where automated methods are inefficient. For example, while names of drugs or diseases can be matched accurately with scripts, annotations of doctors' notes or descriptive text are more challenging to automate due to their varied content. This aspect of annotation could benefit from enhanced text approximation techniques to improve scalability.

The annotations created and the complete annotations of the MIMIC III dataset could serve as a basis for training and benchmarking automated annotation tools. Such tools would simplify administrative tasks in healthcare, potentially making a significant impact, especially in regions still reliant on partial or fully physical records. Automating routine data transfer processes could improve efficiency in handling complex patient histories, thereby reducing costs and enhancing healthcare accessibility.

5 Conclusion

This research developed a comprehensive annotation approach for the MIMIC III Demo dataset, successfully mapping tabular data elements to standard medical ontologies. The records of a single patient were fully annotated with column type, column property, and cell entity annotations. Given the complexities associated with medical records, these annotations can be expanded to encompass the entire dataset or utilized as they currently are. This work sets the foundation for future applications in training and testing automated annotation tools, contributing to advancements in the automation of medical data handling.

6 Implementation and Data Availability

All data and code pertinent to the analysis presented in this work are available in the GitHub repository: `https://github.com/TohaRhymes/auto_anno_um`.

References

[1] Abhisek Sharma, Sumit Dalal, and Sarika Jain. "SemInt at SemTab 2022". In: *SemTab@ISWC*. 2022. URL: https://api.semanticscholar.org/CorpusID:255943679.

[2] Shervin Mehryar and Remzi Celebi. "Semantic Annotation of Tabular Data for Machine-to-Machine Interoperability via Neuro-Symbolic Anchoring". In: *SemTab'23: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching 2023, co-located with the 22nd International Semantic Web Conference (ISWC)*. Corresponding author: Shervin Mehryar (shervin.mehryar@maastrichtuniversity.nl). CEUR Workshop Proceedings. Athens, Greece, Nov. 2023. URL: http://ceur-ws.org/Vol-3557/paper5.pdf.

[3] Sören Auer et al. "DBpedia: A Nucleus for a Web of Open Data". In: (2007), pp. 722–735. DOI: 10.1007/978-3-540-76298-0_52. URL: https://doi.org/10.1007/978-3-540-76298-0%5C_52.

[4] Denny Vrandečić and Markus Krötzsch. "Wikidata: A Free Collaborative Knowledge Base". In: *Communications of the ACM* 57.10 (2014), pp. 78–85. DOI: 10.1145/2629489.

[5] Enrique Iglesias et al. "SDM-RDFizer: An RML Interpreter for the Efficient Creation of RDF Knowledge Graphs". In: *L3S Research Center Leibniz University of Hannover* (2023).

[6] Bader Aldughayfiq et al. "Capturing Semantic Relationships in Electronic Health Records Using Knowledge Graphs: An Implementation Using MIMIC III Dataset and GraphDB". In: *Healthcare* 11.1762 (2023), pp. 1–25. DOI: 10.3390/healthcare11121762. URL: https://doi.org/10.3390/healthcare11121762.

[7] Sören Auer et al. "DBpedia: A Nucleus for a Web of Open Data". In: *The Semantic Web* (2007), pp. 722–735.

[8] Kewei Lyu et al. "Causal knowledge graph construction and evaluation for clinical decision support of diabetic nephropathy". In: *Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University* (2024).

[9] D. Nicholson and C. Greene. "Constructing knowledge graphs and their biomedical applications". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 1414–1428. DOI: 10.1016/j.csbj.2020.05.017.

[10] Ming Sheng et al. "A data-intensive CDSS platform based on knowledge graph". In: *Proceedings of the Health Information Science: 7th International Conference, HIS 2018*. Cairns, Australia, May 2018, pp. 146–155.

[11] Chaoyu Zhang and Lei Li. "A Review of Medical Decision Supports Based on Knowledge Graph". In: *Data Analysis and Knowledge Discovery* 4.1 (2020), pp. 26–32.

[12] Z. Wu, R. Xue, and M. Shao. "Knowledge graph analysis and visualization of AI technology applied in COVID-19". In: *Environmental Science and Pollution Research* 29 (2022), pp. 26396–26408. DOI: 10.1007/s11356-022-19364-5.

[13] Adrien K Lawrence, Liselotte Selter, and Urs Frey. "SPHN-The Swiss Personalized Health Network Initiative." In: *MIE*. 2020, pp. 1156–1160.

[14] Christophe Gaudet-Blavignac. "Semantic interoperability of clinical data: A multi-dimensional approach". In: (2021). DOI: 10.13097/archive-ouverte/unige:157668.

[15] Antoine Bordes et al. "Translating Embeddings for Modeling Multi-relational Data". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.

[16] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[17] Victor Sanh et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: *arXiv preprint arXiv:1910.01108* (2019).

[18] SNOMED International. *SNOMED CT Browser*. Accessed: 2024-06-27. 2024. URL: `https://browser.ihtsdotools.org/`.

[19] Sunghwan Kim et al. “PubChem 2023 update”. In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D1373–D1380. DOI: `10.1093/nar/gkac956`.