From Tables to Graphs with ClinicoAtlas: Leveraging LLMs to Support Modeling and Mining Knowledge Graphs

Rafael C. G. Conrado[1], **Eduardo M. Campos**[2],
Caetano Traina Jr.[1], **Agma J. M. Traina**[1], **Mirela T. Cazzolato**[1]

[1] Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP) – São Carlos, SP – Brazil

[2]Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP)
Universidade de São Paulo (USP) – Ribeirão Preto, SP – Brazil

{rafaelconrado,educampos}@usp.br, {caetano,agma,mirela}@icmc.usp.br

Abstract. *Given a set of Electronic Health Records organized as tabular data, how can we model and evaluate relationships between various concepts related to health conditions and treatments? Representing EHRs as Graphs allows a much more intuitive interpretation of relationships between concepts and permits the mining of node features, generating relevant metrics for hospital data analysis. However, these metrics are often generic and difficult to interpret. In this context, Large Language Models can assist in constructing these graphs by defining the semantic relationships between concept pairs and attributing contextual meaning to the extracted features. The proposed ClinicoAtlas tool integrates LLMs into the construction of knowledge graphs, facilitating the extraction and interpretation of meaningful information for healthcare management.*

1. Introduction

Electronic Health Records (EHRs) are responsible for organizing patient information collected from various sources, including hospitals, laboratories, and clinics. As a central component of health-management systems, EHRs store data such as medical exams, drugs, procedures, conditions, and more.

Consequently, the exploratory data analysis of the EHRs enables more precise decision-making in the health institutions context. In recent years, a promising approach has been to model these data as Knowledge Graphs, which allows for tasks such as information extraction and inference [Murali et al. 2023]. Graph Mining, for instance, is a method utilized across many fields such as bioinformatics, social networks, and chemical reactions [Rehman et al. 2012], as it enables the extraction of metrics like Degree Centrality, Graph Density, and even node features such as In and Out-Degree. However, metrics like node features, despite their relevance, are often generic and challenging to interpret, particularly when applied to data from healthcare institutions.

Large Language Models (LLMs) are capable of providing human-like responses, making them useful in a variety of scenarios. For instance, LLMs have been applied in the construction and reasoning of Knowledge Graphs [Zhu et al. 2024], providing the semantical explanation of the graph connections. This automatic task can reduce the effort for manually providing edge labels between concepts, which could be laborious and time-consuming for domain specialists.

In this work, we propose a tool that leverages LLMs to construct Knowledge Graphs and define the semantic relationships between their concepts, which will be instrumental for feature extraction from graphs. The source code is available in a public repository[1]

2. Background

In our approach, we model two distinct types of graphs: The first one is what we call the 'Conceptual Graph' – which consists of a Knowledge Graph (KG) used to define semantics between different medical concepts extracted from EHRs. This is the modeling based on the attribute concepts (the 'intensional database'). The second one is what we call the 'ClinicoGraph', in which we perform pairwise mining of node features based on the previously defined semantics but using the attribute values as nodes (the 'extensional database'). Next, we present the main concepts related to our proposal, including graph modeling and Large Language Models (LLMs).

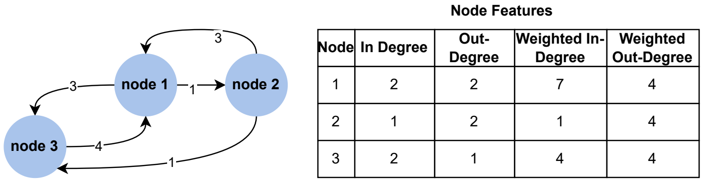
'Conceptual Graph' Modeling. The Conceptual Graph is modeled as a directed and weighted graph, with the format $G = (U, V, E, W)$, where $U = (u_1, ..., u_n)$ is the set of source vertices and $V = (v_1, ..., v_m)$ is the set of destination vertices. Each vertex represents a concept extracted from the column names of the tables in our database. An edge $e \in E$ connects a pair $(u_i, v_j), u_i \in U, v_j \in V$, indicating a relationship between two concepts. Each edge is labeled with the name of the relationship and has an associated weight $w \in (W)$.

The 'ClinicoGraph' Modeling. The ClinicoGraph is a bipartite, directed, weighted, and time-evolving, represented as $G' = (U, V, E, W, T)$, where $U = (u_1, ..., u_n)$ and $V = (v_1, ..., v_n)$ are the sets of all vertices, each corresponding to a value extracted from a table attribute in the database. E is the set of directed edges, representing relationships between $(u_i, v_j), u_i \in U, v_j \in V$. Each edge has an associated weight $w \in W$, which may represent either the number of occurrences of the relationship between attribute values or a specific attribute value linked to that edge. T is the optional set of timestamps, indicating the time of events that linked a pair of attribute values.

'ClinicoGraph' Mining. In this work, we focus on the mining of node features from bipartite graphs, as introduced in a previous work [Conrado et al. 2024]. The chosen metrics include In and Out-Degree, which represent the number of unique vertices connected to a given vertex via incoming and outgoing edges, respectively; Weighted In and Out-Degree, which correspond to the sum of weights of incoming and outgoing edges from set W; Inter-Arrival-Times (IAT), that computes time intervals between successive arrival of edges, using associated timestamps from set T, and Core Number (coreness), which indicates how well-connected a vertex is within the overall structure of the graph.

LLMs. Large Language Models (LLMs) are Artificial Intelligence Models trained in extensive amounts of textual data, allowing the generation of responses, summarization, paraphrasing, and text translation with a level of proficiency comparable to human

[1] ClinicoAtlas is open-sourced at `https://github.com/RafaelCGConrado/clinicoAtlas`.



Node Features

Node	In Degree	Out-Degree	Weighted In-Degree	Weighted Out-Degree
1	2	2	7	4
2	1	2	1	4
3	2	1	4	4

Figure 1. Example of Node Features in a directed, weighted graph

capabilities [Clusmann et al. 2023]. The quality of responses generated by LLMs can be enhanced by incorporating more detailed context into the prompt - in other words, the input given to the model -. This approach improves precision and alignment with the user's request. Moreover, different prompting techniques significantly impact the model's performance, such as zero-shot prompting, where no example is provided, and few-shot-prompting, in which the model receives a small number of examples before generating responses.

LLMs and KGs. The KG Engineering can be a laborious, and prone to errors process. In recent literature, LLMs have been applied to assist the efficient construction of KGs. In [Meyer et al. 2024], the authors conduct experiments with the model ChatGPT to explore its potential to support KG Engineering. In [Yang et al. 2025], the authors classify and evaluate three approaches of combining LLMs and KGs, and also describe essential metrics for assessing the performance of these integrations.

3. ClinicoAtlas: The proposed method

In this work, we propose ClinicoAtlas, a tool developed to understand the semantics of relationships between distinct concepts and node features supported by LLMs. The defined semantics are then applied to the mining of the ClinicoGraph, a bipartite graph modeled from EHR databases. The approach is built with two layers:

(i) Semantic Knowledge Layer. The first step involves specifying the pairs of concepts to be modeled in the conceptual graph. Each concept corresponds to a column (attribute) in a database table. Figure 2 shows the general pipeline. Possible concept pairs are derived from the set of all primary/foreign keys defined in the (Relational) database. Once all possible combinations have been collected, the user can select specific concept pairs of interest for the semantic relationship definition. Each pair is inserted into a predefined prompt, which is then submitted as input to a LLM. The model's output is the proposed label for the semantic relationship. A domain expert in the medical field may oversee this process and modify the prompt if necessary. If the result is deemed appropriate, the triple `<concept1, concept2, label>` is added to the conceptual graph.

The next step is to obtain the semantic interpretation of node feature metrics for each defined concept pair. A second predefined prompt is sent to the LLM, containing the concept pair and the specified feature to be interpreted. The considered node features are In and Out-Degree, Weighted-In and Weighted-Out-Degree, Inter-Arrival-Time (IAT), and Core Number. For each of these metrics, we compute the mean, median, standard deviation, and IQR values. The semantic relationship labels and node feature interpretations are stored in the Knowledge Dictionary, which is the output of this layer.

(ii) Mining Layer. The user can submit SQL Queries on the EHR database. Fig-

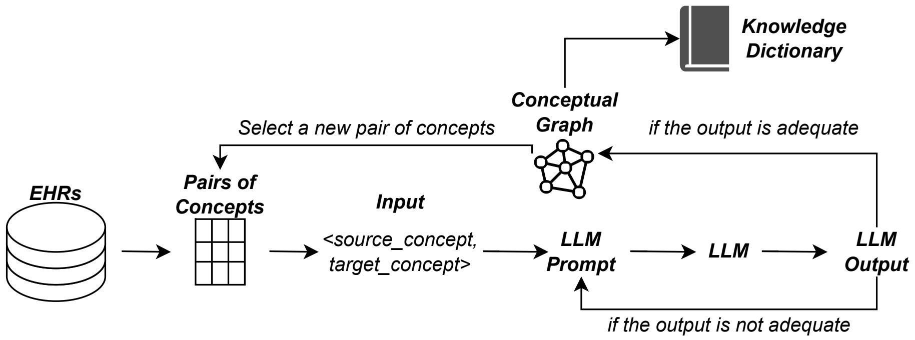

Knowledge Dictionary
Conceptual Graph
Select a new pair of concepts
if the output is adequate
EHRs
Pairs of Concepts
Input
<source_concept, target_concept>
LLM Prompt
LLM
LLM Output
if the output is not adequate

Figure 2. The Semantic Knowledge Layer

ure 3 illustrates this step. From the resulting table, they must select two attributes to proceed with the ClinicoGraph modeling, where ClinicoAtlas performs the mining of node features. The semantics for the selected concept pair, stored in the Knowledge Dictionary from the previous step, are then used to present the final results generated by the graph mining to the user, in an interpretable and context-aware manner.

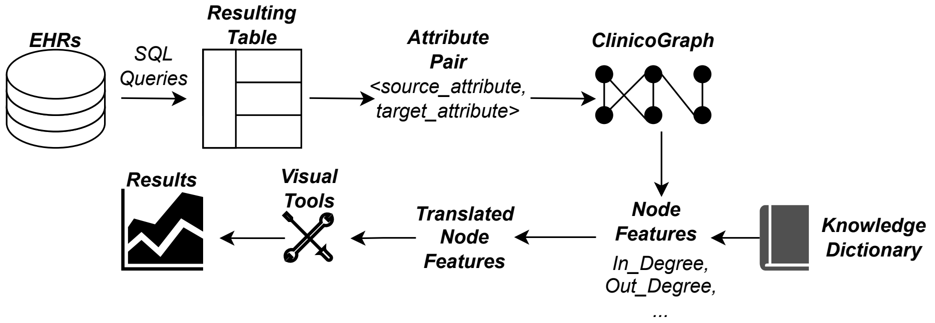


Figure 3. The Mining Layer

4. Results

In this section we show an example to demonstrate the functionality of the ClinicoAtlas. We ran the experiments over a dataset described in [de Lima et al. 2019], with two decades of hospital data.

Figure 4 shows a directed edge type,used to illustrate the tool running of that dataset. Within the Semantic Knowledge Layer, we selected from the *drug_exposure* table the concepts *drug_concept_id*, representing a drug in the database, and *person_id*, which denotes a patient. As a result of the first prompt, we obtained the label *"is administered to"*. We then proceeded to define the names for the node features In and Out Degree and Core Number in the specific context between drug and patient. We received *"Number of patients who received the drug", "Number of drugs the patient received" and "Measure of Centrality of the graph"*, respectively. The generated label and translated-features were all stored in the Knowledge Dictionary.

Once the semantics were defined, we proceeded within the Mining Layer. For our

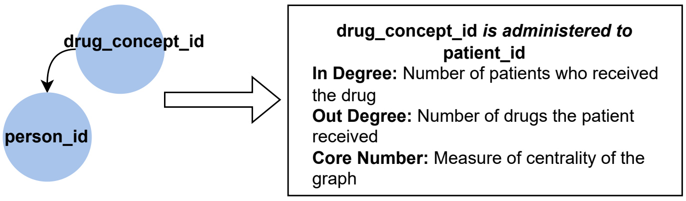

drug_concept_id
person_id
drug_concept_id *is administered to* patient_id
In Degree: Number of patients who received the drug
Out Degree: Number of drugs the patient received
Core Number: Measure of centrality of the graph

Figure 4. Result of the Semantic Knowledge Layer

ClinicoGraph Modeling, the attributes *drug_concept_id* and *person_id* were maintained. After the graph was constructed, the node features from Table 1 were extracted.

Table 1. Node feature results from the Mining Layer.

Nodes	Number of patients who received the drug	Number of drugs the patient received	Measure of Centrality of the Graph
CITRATO DE SILDENAFILA 20MG	1	0	1
BOSENTANA 62,5 COMPRIMIDO	2	0	1
person_568713	0	2	1
person_6728169	0	1	1

Figure 5 shows the ClinicoGraph, which the user can also visualize and interact with. The node size is the corresponding degree, and the color indicates if the node is from the left (source) or right (destination) side of the bipartite modeling.

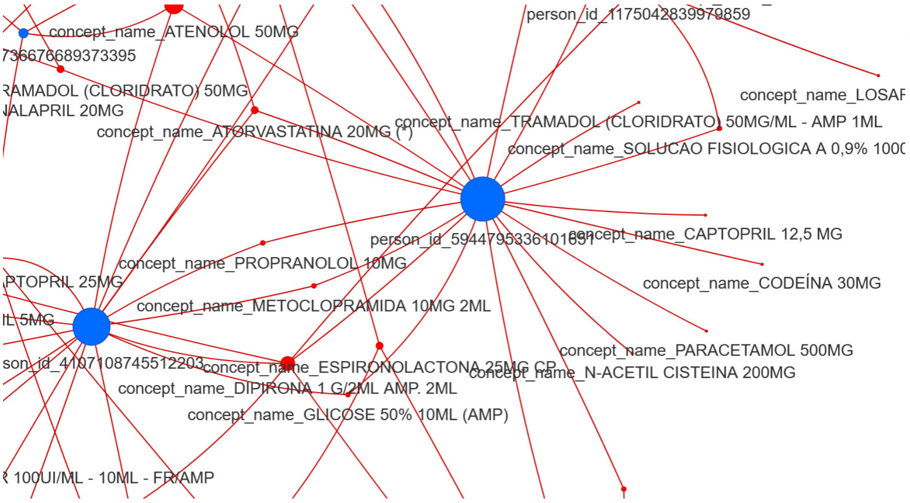


Figure 5. ClinicoGraph visualization

5. Conclusion

In this work, we present ClinicoAtlas, a tool developed to automatically identify semantics between concepts in EHRs databases, and integrate them into Graph Mining tasks. Our main goal is to enhance the interpretability of feature-based analysis, providing a more meaningful and context-specific analysis. The preliminary results show the usefulness of our approach.

The proposed tool ClinicoAtlas can be applied to other domains as well. For this, the predefined prompts need to be adjusted according to the modeled problem, but the other functionalities should work seamlessly.

Acknowledgments

We thank the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001, the São Paulo Research Foundation (FAPESP) – grants #2016/17078-0, #2023/14390-7, #2024/13328-9, the Brazilian National Council for Scientific and Technological Development (CNPq), PIBIC-CNPq – grant #125355/2024-6, Programa Unificado de Bolsas de Estudo para Apoio e Formação de Estudantes de Graduação (PUB-USP) – grant #2075, and the University of São Paulo (PRPI Ordinance No. 1032) – grant #207.

References

Clusmann, J., Kolbinger, F., Muti, H., Carrero, Z., Eckardt, J.-N., Laleh, N., Löffler, C., Schwarzkopf, S.-C., Unger, M., Veldhuizen, G., Wagner, S., and Kather, J. (2023). The future landscape of large language models in medicine. *Communications medicine*, 3:141. DOI: 10.1038/s43856-023-00370-1.

Conrado, R., Gutierrez, M., Jr., C. T., Traina, A., and Cazzolato, M. (2024). Combining semantic graph features and a common data model to exploit the interoperability of patient databases. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 701–707, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2024.243153.

de Lima, D. M. et al. (2019). Transforming two decades of ePR data to OMOP CDM for clinical research. In *MEDINFO 2019*, volume 264, pages 233–237. IOS Press. DOI: 10.3233/SHTI190218.

Meyer, L.-P., Stadler, C., Frey, J., Radtke, N., Junghanns, K., Meissner, R., Dziwis, G., Bulert, K., and Martin, M. (2024). *LLM-assisted Knowledge Graph Engineering: Experiments with ChatGPT*. Springer Fachmedien Wiesbaden. DOI: 10.1007/978-3-658-43705-3_8.

Murali, L., Gopakumar, G., Viswanathan, D. M., and Nedungadi, P. (2023). Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study. *Journal of Biomedical Informatics*, 143:104403. DOI: https://doi.org/10.1016/j.jbi.2023.104403.

Rehman, S. U., Khan, A. U., and Fong, S. (2012). Graph mining: A survey of graph mining techniques. In *Seventh International Conference on Digital Information Management (ICDIM 2012)*, pages 88–92. DOI: 10.1109/ICDIM.2012.6360146.

Yang, W., Some, L., Bain, M., and Kang, B. (2025). A comprehensive survey on integrating large language models with knowledge-based methods. *Knowledge-Based Systems*, 318:113503. DOI: 10.1016/j.knosys.2025.113503.

Zhu, Y., Wang, X., Chen, J., Qiao, S., Ou, Y., Yao, Y., Deng, S., Chen, H., and Zhang, N. (2024). Llms for knowledge graph construction and reasoning: recent capabilities and future opportunities. *World Wide Web*, 27(5):58. DOI: 10.1007/s11280-024-01297-w.