Combining Semantic Graph Features and a Common Data Model to Exploit the Interoperability of Patient Databases

Rafael C.G. Conrado[1], Marco A. Gutierrez[2], Caetano Traina Jr.[1], Agma J.M. Traina[1], Mirela T. Cazzolato[3]

[1] Institute of Mathematics and Computer Sciences (ICMC)
University of São Paulo (USP) - São Carlos, Brazil

[2]Institute of Heart (InCor)
Clinical Hospital of Faculty of Medicine (FMUSP) - São Paulo, Brazil

[3]Faculty of Philosophy, Sciences and Letters at Ribeirão Preto (FFCLRP)
University of São Paulo - Ribeirão Preto, Brazil

{rafaelconrado, mirelac}@usp.br

Abstract. *Given a set of Electronic Health Records (EHRs), how can we semantically model the available concepts and provide tools for data analysis? EHRs following a common data model (CDM) usually provide meaningful organization and vocabulary to health-related databases, prompting data interoperability. However, hidden relationships among attributes within the CDM bring the need for CDM-tailored analysis tools regarding exploratory tasks. We propose GraFOCAL for analyzing CDM-based databases considering semantic graph features. GraFOCAL combines pairs of attributes with semantic descriptions in graph edges and node features. Preliminary results show the usefulness of GraFOCAL's features and visual tools in spotting findings in a real-world dataset. In future work, we aim to extend the proposed approach with automatic knowledge inference for the semantic linkage between variables.*

1. Introduction

Electronic Health Records (EHRs) are at the core of a healthcare management system, since they organize patients' data from health institutions, such as hospitals, clinics, and labs. EHRs contain data from different aspects related to everyday activities, such as patient information, exams, drugs, administrative actions, and staff management [Andrade and Medeiros 2023]. Managing, integrating, and processing a large amount of information among different institutions is a critical challenge to provide interoperability. Concretely, EHRs used in patient care across different institutions often have incompatible structures and terminology. This complexity inhibits our ability to process and analyze the information, posing a significant load in effective decision-making. A promising solution for this scenario is employing a Common Data Model (CDM) such as OMOP (the Observational Medical Outcomes Partnership) [OHDSI 2024]. However, seamlessly processing the distinct semantic relationships inherent to the model is far from trivial.

Leveraging semantic graphs in this context can support integrating the heterogeneous concepts of EHRs and strengthen the data analysis for decision-making [Yang et al. 2024]. As CDM provides significant and standardized values for the attributes, exploring correlations between attribute pairs potentially becomes much more

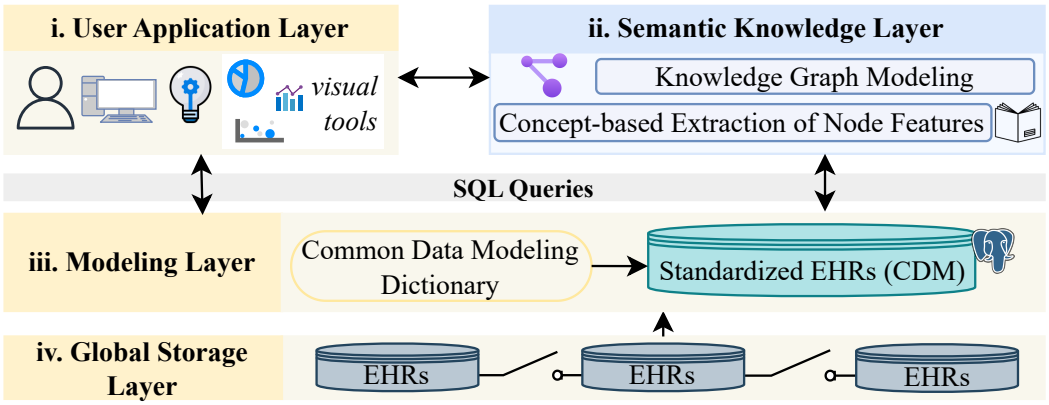

Figure 1. *GraFOCAL* overview: *Layer i* handles the user interface for query execution, graph modeling with attribute pairs, and visualizing results; *Layer ii* translates pairs of variables into semantic relationships and extracts meaningful features; *Layer iii* handles EHRs modeled into a CDM (in this work, we adopt the OMOP-CDM); *Layer iv* stores EHRs repositories, with data retrieved from many sources.

useful. Such modeling takes advantage of standardized values by adopting a CDM, combining them with other attributes, and providing insights regarding the observed patterns. Once modeled, features can be extracted from the graph structure and the obtained information is explored through visualization metaphors [Cazzolato et al. 2023]. The unsupervised approach allows users to perform exploratory data analysis (EDA) over EHRs.

Analyzing features extracted from the graph is meaningful, but it also poses challenges regarding the semantic meaning of the relationships. Many works from the literature have already explored graph information for data mining [Xiao et al. 2023, de Souza et al. 2022], IoT [da Costa et al. 2022], anomaly detection [Gupta et al. 2018], and healthcare analysis [Wang et al. 2024]. However, given their technical meaning, the approaches provide generic measures that, although meaningful, are hard to understand.

In this work, we address those challenges by modeling and analyzing semantic graphs of EHRs. The proposed *GraFOCAL* is a method that combines semantic relationships, node features, and visualization tools for EHR data analysis. Figure 1 illustrates the *GraFOCAL* structure and modules. Its main contributions are in layers i and mostly ii, since the latter provides a graph modeling with semantic knowledge from a CDM-based dataset. The graph modeling and source code are available in a public repository[1]. In summary, the main contributions of this work are: (i) Employing semantic graph concepts to model the relationships among OMOP-CDM elements; (ii) employing curated features extracted from the graph for data analysis; (iii) tailoring a set of visual tools to combine the extracted features for the analysis task, improving explainability; (iv) evaluating the tool to analyze a real-world EHR dataset.

2. Background

Graph Modeling and Mining. We model our approach as a **weighted, directed time-evolving graph**. The graph has the form $G = (U, V, E, W\{, T\})$, where $U = \{u_1, ..., u_s\}$,

[1] *GraFOCAL* is open-sourced at `github.com/rafaelCGConrado/grafocal`.

$V = \{v_1, ..., v_t\}$ are sets of vertices such that $U \cup V$ has all the vertices from G. **Each vertex** is a **value from a table attribute** in the database. Set E represents all the graph edges, defining pairwise relationships between attribute values (vertices) we want to link.

Each edge has a direction (from a source to a destination node), and a label to indicate the semantics of the relationship between attributes. Set W is the weights of each edge in E. In our modeling, the weight can correspond to values of a numeric attribute linked by the specific edge. The user can also set the edge weight as the number of edges between the modeled attribute values. For example, suppose we have a pair of attributes 'exam' and 'result'. We can have the edge between vertices 'exam=Covid' and 'result=True' appearing y times. Thus, the edge weight between these two nodes is y, corresponding to the edge frequency of these two concepts. Finally, the set of timestamps T is optional, and indicates the time of the event that links the pair of attribute values.

Node Feature Extraction. We can extract many features from the graph nodes, edges, and the topology [Wang et al. 2024]. **We focus on node features** since they trivially summarize the needed information, are easy to understand, and fast to extract. Among the most explored graph features are the degree, weighted degree, inter-arrival times (IAT), and core [Cazzolato et al. 2023]. The degree counts the number of unique nodes connected to a node. The weighted degree sums the weights of edges linked to that node. IAT features compute the interval between each edge when we have the set of timestamps modeled as well. All features (except core) have 'plain' values (when we ignore the edge direction), 'in' values for incoming edges, and 'out' values for outgoing edges. We can also compute statistic measurements from in/out IAT and in/out weighted degrees, such as the average, the standard deviation, and inter-quantile ranges (IQRs). Finally, the core number is obtained from the k-core graphs and informs how well-connected the node is.

The OMOP-CDM. OMOP was created as a public-private partnership between pharmaceutical companies and the American National Institutes of Health, now with global reach. OMOP established a program to advance the field of medical active-monitoring products of health-related data [Stang et al. 2010]. OMOP-CDM is a mechanism developed to standardize structure, content, and semantics from observational data, establishing standardized vocabularies with different types of information stemming from various sources, fostering collaboration between different institutions [Overhage et al. 2011]. The CDM is optimized to identify populations of patients with specific medical interventions and their respective outcomes. It is 'Person-Centric': every Event table is related to the 'PERSON' table and platform-independent, meaning that every data type is defined in a generic form following the ANSI SQL standard [OHDSI 2024].

3. Related Work

We cover studies from literature related to the analysis of EHRs, graph features, and explainability. Table 1 summarizes the related work regarding the following criteria: ***Graph features***, meaning the approach works with features extracted from the graph structure, nodes, or edges; ***Works with EHRs (or data from the medical domain)***, when the main goal of the work is to support health-related problems or exploratory data analysis (EDA); ***CDM-oriented semantics***, when the explored EHRs are modeled following a CDM and provides semantic tools for decision-making; as well as ***Visual tools***, for analyses that take advantage of visual representations to improve explainability.

In [Nouri et al. 2021], the authors propose a visual tool for EHRs (Visemure), with statistical and machine learning approaches. They combine different information in the data for EDA of preprocessed datasets. In [Cazzolato et al. 2023] the authors modeled EHRs as weighted and directed graphs, extracting node features for data mining. They introduced GraF-EDA, which combines graph features with visualization tools for EDA. The visual tools include 2-D scatter plots, n-D scatter matrices, interactive lasso tools, graph visualizations (with spring layout), and hex-bins. In [Xiao et al. 2023], the authors propose FHIR-Ontop-OMOP, an ontology for OMOP-CDM to build knowledge graphs. They propose a component to query standardized data using standard SQL, 'translating' the query information to the OMOP-CDM concepts.

Several other works provided graph mining tools for other domains, such as EDA and fraud detection, but they did not address health-related problems and their underlying semantics [Gupta et al. 2018, Fidalgo et al. 2022].

Method	Graph Features	EHRs	CDM-oriented semantics	Visual Tools
Visemure	✗	✓	✗	✓
GraF-EDA	✓	✓	✗	✗
FHIR-Ontop-OMOP	✗	✓	✓	✗
Proposal: *GraFOCAL*	✓	✓	✓	✓

Table 1. *GraFOCAL matches all the required specifications:* comparison between our proposal and existing works from the literature.

Considering the related works studies and gaps identified, we propose *GraFOCAL* for time-evolving graph mining using node features and visualization tools. Our method focuses on modeling EHR concepts as a semantic graph, bringing meaningful reasoning to the analysis. We detail *GraFOCAL* next.

4. *GraFOCAL*: The Proposed Method

In this work, we propose *GraFOCAL* (Graph Features for Common Data Models) for modeling semantic relationships in weighted, directed graphs built over attributes of databases following a Common Data Model (CDM). Specifically, in this work, we focus on the OMOP-CDM, but the approach can be easily adapted for other models as well. The approach was designed with four layers, described as follows.

(i) User application: The user can pose SQL queries on the available EHRs from *layer iii*. The user must select two attributes in the resulting table through such queries to proceed with the graph modeling. After the semantic modeling and feature extraction of *layer ii*, the user can visualize the semantic patterns with carefully chosen visual tools.

(ii) Semantic Knowledge: This is the primary step of our proposal. Given a pair of attributes, *GraFOCAL* models the data as weighted, directed graphs and extracts meaningful features from the nodes.

(iii) Modeling: This layer maintains EHRs already preprocessed and translated/converted to a Common Data Model (CDM). Following the pattern-defined modeling, all SQL statements are posed over standardized EHRs.

(iv) Global Storage: In this layer, we can have databases from many institutions following different patterns and namings. Converting the database from layer iv to layer iii

requires considering the proposed modeling of the adopted CDM and selected vocabulary. This conversion must also rely on the specialist's knowledge since it involves many medical concepts inherent in the data domain.

The Semantic Knowledge Layer. *GraFOCAL* models attributes as vertices and links them using semantic relationships defined for the specific CDM adopted. We adopt OMOP-CDM and define the set of relationships among the different concepts described in the modeling. Figure 2 shows a few selected nodes of the modeled graph and the corresponding semantic relationships. We also define the meaning of features extracted from the semantic graph, giving them the proper nomenclature to improve explainability. An example of a semantic rule and corresponding features produced in our modeling is:

`person.person_id` *is exposed to* `drug_exposure.drug_concept_id`
The *in-degree* gives the number of distinct drugs the person has been exposed to. The *core number* of `person_id` informs how well-connected the drugs the person takes are with other patients as well. The *out-average IAT* gives the interval of time the patient takes a given drug.

'Drug' is the term used in OMOP-CDM for medicines and other remedies used to treat someone. The graph features extracted are those described in Section 2 (*Node Feature Extraction*). The full semantic labels for relationships (node edges) and features are stored as a table in the dataset, and we make the `CREATE TABLE` script with all information available in our repository[1], so readers can extend the approach to other CDMs. After feature extraction, *GraFOCAL* translates the feature meaning using the reference table and shows it to the user with visual tools.

The User Application Layer *GraFOCAL* provides a user interface to pose queries over the EHRs, runs the graph modeling and feature extraction over selected attributes, and produces a visualization of the results. The interface is developed using the Streamlit framework (`https://streamlit.io`), and the visual tools are the hexbin, interactive 1-D and n-D scatter plots, as well as graph visualization with the spring layout. With these tools, our method can visualize million-scale graphs, combining two or more features at a time, and using log-scales and colors to help illustrate large ranges of values.

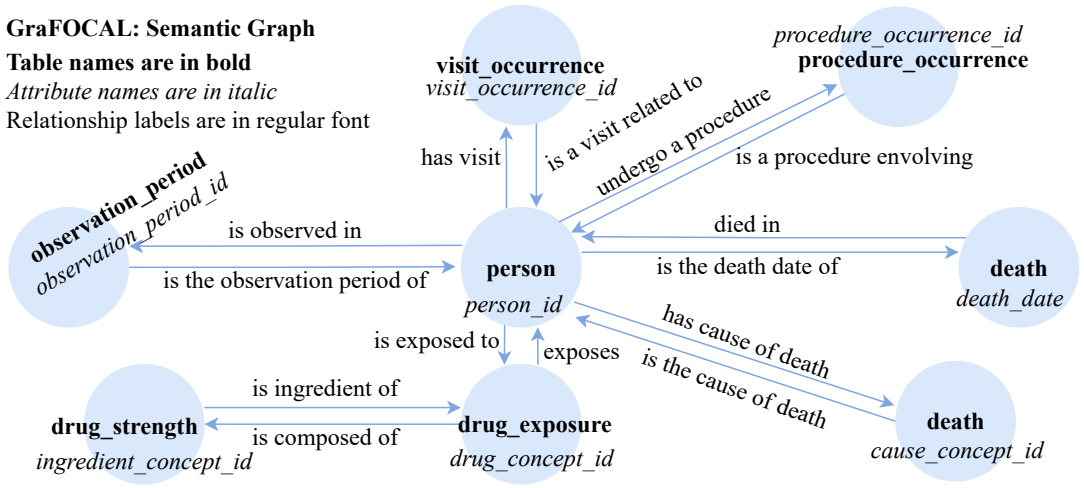

Figure 2. Snapshot of the semantic graph built for the OMOP-CDM modeling. Each pair of connected attributes has two relationship labels: one for incoming and one for outgoing edges.

All plots are shown using the Python libraries Matplotlib, Plotly, NetworkX, and PyVis.

5. Preliminary Results

Dataset. We ran the experiments over the dataset provided in [de Lima et al. 2019], with two decades of hospital anonymized data converted to OMOP-CDM. The dataset provides data from $94,603$ patients, with more than 30 million tuples in 40 tables, such as *concept, condition_occurrence, person, and procedure_occurrence*.

Analysis case: Figure 3 exemplifies the modeling two concepts: `procedures` (*e.g.,* exams) taken by patients with different `conditions`. We randomly selected $1,000$ tuples and modeled the links between the conditions and procedures of each patient on a given procedure date. The graph weight is the edge frequency between the pairs of nodes. We can notice that the different conditions share many procedures. The procedure 'cardiology consultation' was required for the most distinct conditions (10) and is among the most well-connected procedures in the sample. The procedures with the highest frequency were 'creatinine', 'platelet count', and 'complete blood count' ($60\times$ each). The 'UTI' condition required 85 distinct procedures, 759 times in total. The computational cost for feature extraction is linear to the number of edges, except for the core number which requires a constant number of data passes.

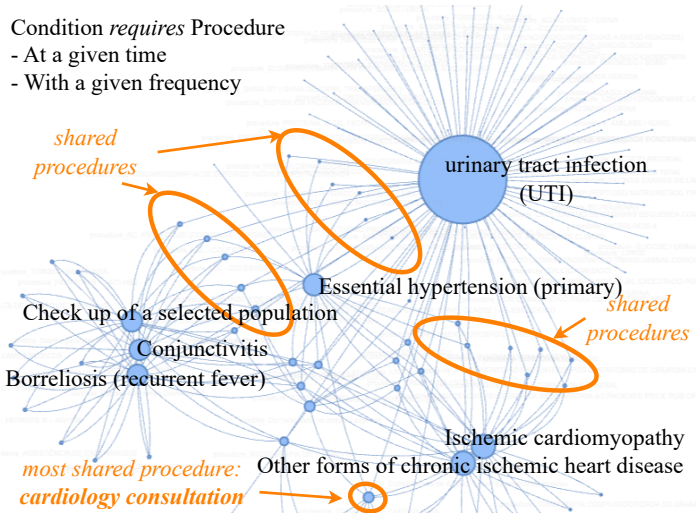

Figure 3. *GraFOCAL* at work.

6. Conclusion

We proposed *GraFOCAL* for exploring semantic graph features and the OMOP-CDM to improve the analysis of EHRs. *GraFOCAL* composes a graph with attributes defined by OMOP and adds a knowledge layer over the concepts to improve pattern explainability. The method extracts node features from the concepts in the graph and shows the observed patterns using visual tools. The preliminary results show the usefulness of our approach.

This work is one of the first steps towards providing a semantic layer over CDM-based databases for graph mining. We consider extending our work by automatically inferring the semantic relationships between attributes using approaches such as language models and employing similarity queries to compare different concepts and cases.

Acknowledgment

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 No. 2016/17078-0, 2020/11258-2), CNPq and Programa Unificado de Bolsas de Estudo para Apoio e Formação de Estudantes de Graduação (PUB-USP) - grant #2912.

References

Andrade, M. J. and Medeiros, C. B. (2023). Linking heterogeneous health data sources in brazil centered on drug leaflet processing. In *SBBD 2023*, pages 366–371. SBC. DOI: 10.5753/sbbd.2023.233356.

Cazzolato, M. T. et al. (2023). Exploratory data analysis in electronic health records graphs: Intuitive features and visualization tools. In *CBMS 2023*, pages 117–122. IEEE. DOI: 10.1109/CBMS58004.2023.00202.

da Costa, F. J. et al. (2022). Dikw4iot: Uma abordagem baseada na hierarquia DIKW para a construção de grafos de conhecimento para integração de dados de iot. In *SBBD 2022*, pages 190–202. SBC. DOI: 10.5753/sbbd.2022.224648.

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.

de Souza, E. M. F. et al. (2022). Visualização interativa da evolução de grafos de conhecimento. In *SBBD 2022*, pages 343–354. SBC. DOI: 10.5753/sbbd.2022.224301.

Fidalgo, P. et al. (2022). Star-bridge: a topological multidimensional subgraph analysis to detect fraudulent nodes and rings in telecom networks. In *Big Data 2022*, pages 2239–2242. DOI: 10.1109/BigData55660.2022.10020714.

Gupta, N. et al. (2018). Beyond outlier detection: Lookout for pictorial explanation. In *ECML PKDD 2018*, volume 11051 of *LNCS*, pages 122–138. Springer. DOI: 10.1007/978-3-030-10925-7_8.

Nouri, M. et al. (2021). VISEMURE: A visual analytics system for making sense of multimorbidity using electronic medical record data. *J. Data*, 6(8):85. DOI: 10.3390/DATA6080085.

OHDSI (2024). The Book of OHDSI — observational health data sciences and informatics. `https://ohdsi.github.io/TheBookOfOhdsi/`. Last accessed in 27-06-2024.

Overhage, J. M. et al. (2011). Validation of a common data model for active safety surveillance research. In *Journal JAMIA*, volume 19, pages 54–60. DOI: 10.1136/amiajnl-2011-000376.

Stang, P. et al. (2010). Advancing the science for active surveillance: Rationale and design for the observational medical outcomes partnership. In *Annals of internal medicine*, volume 153, pages 600–6. DOI: 10.1059/0003-4819-153-9-201011020-00010.

Wang, Y., Peng, Y., and Guo, J. (2024). Enhancing knowledge graph embedding with structure and semantic features. In *Appl. Intell.*, volume 54, pages 2900–2914. DOI: 10.1007/S10489-024-05315-2.

Xiao, G. et al. (2023). FHIR-Ontop-OMOP: Querying OMOP clinical databases as fhir-compliant clinical knowledge graphs. volume 3415 of *CEUR Workshop*, pages 165–166. CEUR-WS.org. DOI: 10.1016/j.jbi.2022.104201.

Yang, P. et al. (2024). LMKG: A large-scale and multi-source medical knowledge graph for intelligent medicine applications. *Knowl. Based Syst.*, 284:111323. DOI: 10.1016/J.KNOSYS.2023.111323.