Resolution-Alignment-Completion of Tabular Electronic Health Records via Meta-Path Generative Sampling

Shervin Mehryar
Maastricht University, Minderbroedersberg 4-6, 6211 LH Maastricht
shervin.mehryar@maastrichtuniversity.nl

Abstract

The increasing availability of electronic health records (EHR) offers significant opportunities in data-driven healthcare, yet much of this data remains fragmented, semantically inconsistent, or incomplete. These issues are particularly evident in tabular patient records where important contextual information are lacking from the input for effective modeling. In this work, we introduce a system that performs ontology-based entity alignment to resolve and complete tabular data used in real-world clinical units. We transform patient records into a knowledge graph and capture its hidden structures through graph embeddings. We further propose a meta-path sample generation approach for completing the missing information. Our experiments demonstrate the system's ability to augment cardiovascular disease (CVD) data for lab event detection, diagnosis prediction, and drug recommendation, enabling more robust and precise predictive models in clinical decision-making.

1 Introduction

The amount of data stored as electronic health records (EHR) in tabular format has grown significantly in recent years, now including an immense quantity of interactions, events and interconnected information. As such, data integration will play a transformative role in health information systems for the years to come, bridging the gap between research and applications. Existing machine learning paradigms however cannot directly operate on relational data due to the complex structure of interconnected tables. Domain specific algorithms therefore are in need for efficient and robust processing of tabular EHR for use in clinical decision making (Teng et al., 2020).

In recent years, graph representation learning has been proposed as an approach for modeling relational data where rows become nodes, columns form node features, and primary-foreign key links establish edges. To learn their underlying structure, embedding models have been successfully applied to capture hidden hierarchies for downstream clinical tasks, such as comorbidity and readmission prediction (Choi et al., 2020). In (Robinson et al.), entity-level features are extracted and embedded via Graph Neural Networks (GNN) for training a task-specific model by adopting a schema-less approach, modeling relational data as a heterogeneous graph. While schema-less design offers flexibility, it is less suited for integrating external knowledge sources due to the absence of a predefined structure (Yue et al., 2020).

In contrast, a fixed schema can be imposed enabling seamless extension to external knowledge sources which exist in the form of clinical and biomedical ontologies. However, integrating these sources necessitates ontology alignment to resolve semantic ambiguities and maintain coherent representations. In (Hao et al., 2021), a graph representation learning approach is proposed that maps tabular data sources to a domain specific ontology in order to mitigate the presence of ambiguous information. These models continue to suffer from the inherent incompleteness, missing values, and inconsistent codification from legacy systems.

In this work, we propose a robust resolution-alignment-completeness (RAC) system for consolidating tabular EHR into semantically consistent health knowledge graphs, using standard terminologies aligned with medical ontologies. Unlike prior schema-less, graph-based approaches, our fixed schema approach prioritizes structural integration and scalability for enhancing predictive performance by aligning domain-specific knowledge with relational data. Our modular design consists of the following components:

Proceedings of the 4th Table Representation Learning Workshop, pages 200–207

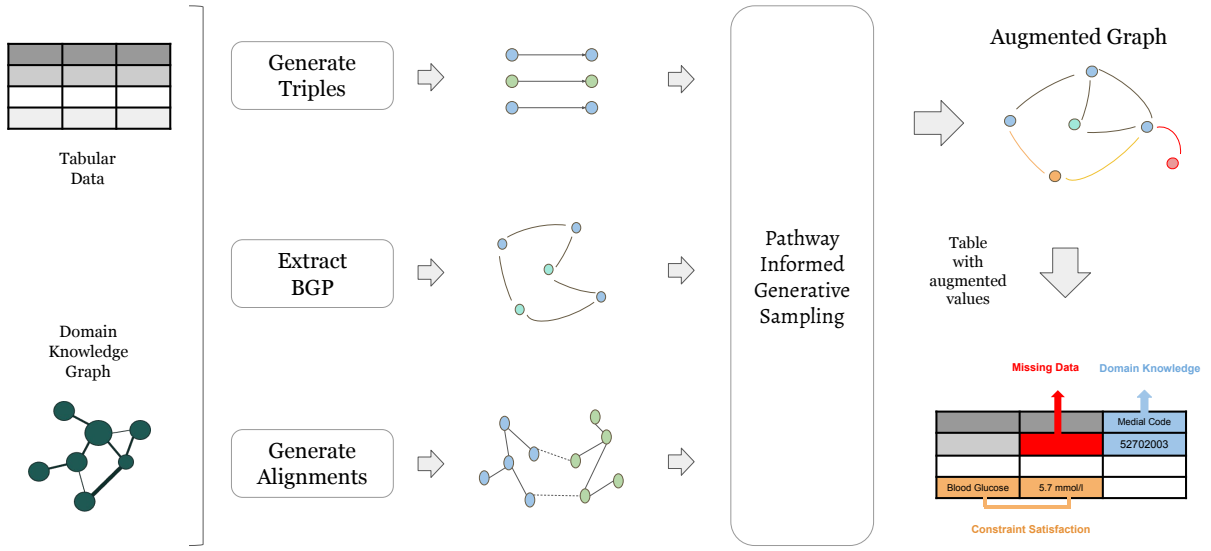


Figure 1: Pathway Informed Generative Sampling and Table Representation through Resolution (R), Alignment (A), and Completeness (C) modules: EHR entities are mapped to Basic Graph Patterns (BPG) of a reference schema, clinical codes are resolved and aligned to SNOMED CT, and meta-path sampling augments representations with missing and task-relevant knowledge.

- **Resolution (R):** In the first module, relevant patient entities are extracted from a relational data source and resolved/mapped via semantically equivalent identifiers and a fixed schema. This module is responsible for identifying and assigning types to data across patient visits using concepts and relations from the fixed schema. Subsequently, the semantically annotated admission records are integrated into a personal health knowledge graph as described in subsection 2.1.

- **Alignment (A):** In the second module, the resulting knowledge graph is transformed and vectorized into a shared embedding space. Through alignment of core concepts with a reference ontology, ambiguous representations are semantically enriched and contextualized, as described in subsection 2.2.

- **Completeness (C):** In the third module, the aligned representations are further enhanced by generating samples along upper ontology concepts (i.e. meta-paths) in the knowledge graph. The samples generate the augmented graph that is used to complete missing information given a prediction task, as described in subsection 2.3.

We use the MIMIC repository for experimentation which contains data associated with distinct hospital admissions concerning adult patients admitted to critical care units (Johnson et al., 2016). In order to map patient relational records, we use the Swiss Personal Health Network Schema (SPHN)[1] and a fine-tuned language model to process the input data. The resulting health knowledge graph is embedded using relational graph neural networks and aligned with the Systematized Nomenclature of Medicine[2] (SNOMED) as domain knowledge graph. We test our framework for three different down-stream clinical tasks, namely lab event detection, diagnosis prediction, and drug recommendation. Our experiments demonstrate the contributions from each component, namely semantic annotation, schema-based entity resolution and domain ontology alignment, to predictive performance using precision, recall, and f1 scores as classification metrics.

2 Method

The meta-path sampling framework proposed for tabular EHR processing in this work, is shown in Figure 1. Related entities from tables are extracted and mapped to the relevant parts represented by Basic Graph Patterns (BPG) in a reference schema. Rows (records) are assigned unique identifiers and instances of the corresponding class and column attributes are mapped to retrieved predicates as triples. Clinical codes (e.g. ICD[3], etc)

[1] https://biomedit.ch/rdf/sphn-schema/sphn
[2] https://www.snomed.org/
[3] https://icd.who.int/

are assigned unique identifiers to resolve their semantically equivalent instances and aligned with a domain-specific ontology, namely SNOMED CT. Lastly, the transformed records are embedded and enriched using meta-pathway informed sampling in order to augment their representations, including missing and domain knowledge, as described in the following subsections. Knowledge represented through this system can ultimately be utilized to complete the input data in tabular format.

2.1 Semantic Annotation

In this section we provide details related to the **Resolution** module, including admission record extraction, semantic annotation, and personal health knowledge graph generation. The existing records from the dataset are grouped according to individual visits and by admission ID into separate tables, thus taking an admission centric view. Subsequently, records from each record are mapped to concepts and relationships from a reference schema using a pre-trained large-language model (LLM) to generate typed entities and properties in form of a personal health knowledge graph. The steps to generate the latter, referred to as **Semantic Annotation**, are shown in Figure 2.

More specifically, the tabular data are transformed into a knowledge graph in this stage in order to enable semantic interoperability required in later stages. To this end, cell values are given a type from a reference schema (column annotation) and cell value pairs are linked through a predicate from the reference schema (property annotation). The mapping from the original relational representations to entities linked with reference predicates can be done using a pretrained LLM (Dasoulas et al., 2023). The output is further processed to produce a PHKG in Resource Description Framework (RDF) format.

The steps for generating the transformed RDF from tabular data using the LLM are summarized in Algorithm 1. The records are processed and mapped around core concepts C from the reference schema (e.g. $C =$ *'Diagnosis'*). Once the type of the concept is identified, the basic graph pattern (BGP) related to C given the record r is retrieved (denoted by C_r). For each record, the LLM is applied in several iterations to retrieve the entity types e for each value and the predicate type p between value pairs using the corresponding BGP.

Algorithm 1 Semantic Annotation with Pretrained Large Language Model

Input: Single patient u records $\mathcal{R}_u$, basic graph patterns for core concepts C, LLM
Output: Personal Health Knowledge Graph $\mathcal{G}_u$ for the patient

Initialize: empty graph $\mathcal{G}_u$
for each record r in $\mathcal{R}_u$ **do**
  $C_r \leftarrow \text{LLM}(r)$ ▷ Determine BGP
  for each pair (c_i, c_j) in r **do**
    $(p_{ij}, e_i, e_j) \leftarrow \text{LLM}(C_r, c_i, c_j)$
    $\mathcal{G}_u \leftarrow \mathcal{G}_u \cup (c_i, p_{ij}, c_j)$ ▷ Add edge
    $\mathcal{G}_u \leftarrow \mathcal{G}_u \cup (c_i, e_i)$ ▷ Add type for c_i
    $\mathcal{G}_u \leftarrow \mathcal{G}_u \cup (c_j, e_j)$ ▷ Add type for c_j
  end for
end for
return $\mathcal{G}_u$

Algorithm 2 Entity Alignment Between PHKG and DSRO

Input: PHKG $\mathcal{G} = \{V, E\}$, DSRO $\mathcal{G}_s = \{V_s, E_s\}$, labeled nodes $V_L = \{v_1, \ldots, v_L\}$, unlabeled nodes $V_U = \{v_{L+1}, \ldots, v_{L+U}\}$, pretrained GCN encoder & decoder {ENC() ,DEC()}, threshold λ
Output: Alignment graph $\mathcal{G}_{\text{align}}$

Fine-tuning Step
Initialize: empty $\mathcal{G}_e$ and $\mathcal{G}_c$
for v_i in V_L **do**
  $\mathcal{G}_e \leftarrow \mathcal{G}_e \cup \{(s, p^+, v_i) \in \mathcal{G}\} \cup \{(v_i, p^+, o) \in \mathcal{G}\}$ ▷ subgraph
  $\mathcal{G}_c \leftarrow \mathcal{G}_c \cup \{(s, p^+, v_i) \in \mathcal{G}_s\} \cup \{(v_i, p^+, o) \in \mathcal{G}_s\}$ ▷ subclass
  # Update Encoder & Decoder
  $\text{ENC}, \text{DEC} \leftarrow \text{DEC}(\text{ENC}(\mathcal{G}_e), \text{ENC}(\mathcal{G}_c))$
end for

# Alignment Step
for v_u in V_U **do**
  for v_s in V_s **do**
    $s \leftarrow \text{DEC}(\langle v_u, v_s \rangle)$ ▷ score
    if $s > \lambda$ **then** ▷ threshold
      $\mathcal{G}_{\text{align}} \leftarrow \mathcal{G}_{\text{align}} \cup \{\langle v_u, v_s \rangle\}$
    end if
  end for
end for
return $\mathcal{G}_{\text{align}}$

The generated types and predicates are added to the personal health knowledge graph $\mathcal{G}$ and returned at the end of the algorithm (Mehryar, 2025).

2.2 Ontological Matching

In this section we provide details related to **Alignment** module, including extracting core concepts, retrieving and encoding the corresponding membership graphs, encoding patient health knowledge graph, and alignment via graph neural network decoding. We rely on a domain specific reference ontology (DSRO) for the alignment task. The coded clinical concepts for each patient are first matched based on their label information with core classes from the reference ontology, non-exhaustively. Subsequently, the target classes are enriched with RDF/s and Web Ontology Language (OWL) hierarchical information, forming a corresponding (subsumption) subgraph. The subsumption graph along with the original personal health knowledge graph are encoded into a shared vector space and further decoded to determine final alignments for **Ontological Matching**, as shown in Figure 3.

More specifically, with Ontological Matching the aim is to align codified information within a personal health knowledge graph (PHKG) according to structural and semantic information of the DSRO required in later stages. To this end, coded information pertaining to core concepts (i.e. diagnosis, procedures, prescriptions etc) are embedded using a graph convolution network (GCN) encoder. The GCN encoder is used to embed the source and target entities, including membership information (i.e. sub- and super-classes). The matching between two sets of encoded representations is established through the GCN decoder trained on labeled information. For the unlabeled entities, the pretrained encoder and decoder are applied to determine matching pairs that score over a pre-specified threshold value.

The steps for generating alignment pairs between the PHKG denoted by $\mathcal{G}$ and the DSRO membership graph denoted by $\mathcal{G}_s$, are summarized in Algorithm 2. The labeled entities $v_i \in V_L$ are first extracted from both sources to produce training graphs $\mathcal{G}_e$ and $\mathcal{G}_c$, respectively. The decoder is fine-tuned on these sets for alignment task and by decreasing the distance between the matching representations (i.e. update step). It is

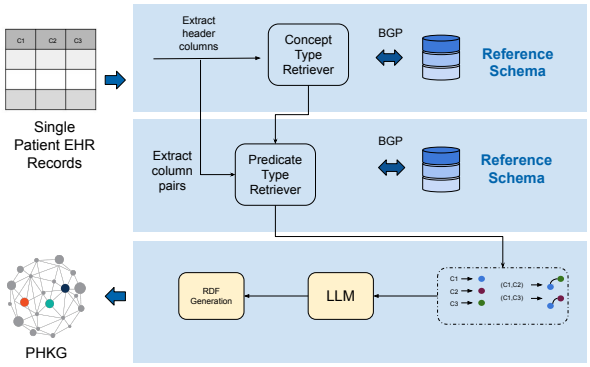


Figure 2: Semantic Annotation using a Large Language Model (LLM), generating health knowledge graph given input electronic health records (EHR) for a single patient, according to basic graph pattern (BPG) extracts from a reference schema.

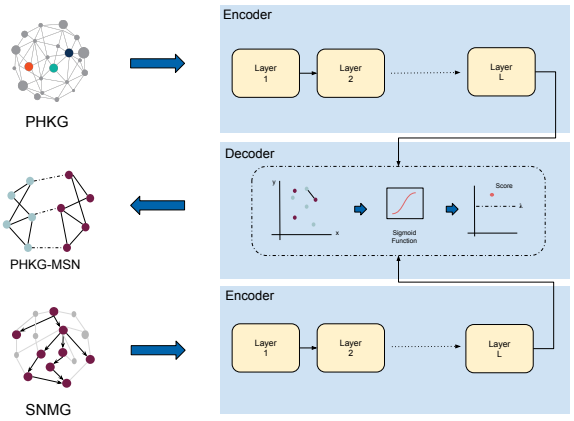


Figure 3: Ontological Matching using a layered (from 1 to L) Graph Neural Network (GNN), generating matches between the entities in a health knowledge graph (PHKG) in alignment with SNOMED CT (SNMG) as domain ontology, to produce the enriched knowledge graph (PHKG-MSN).

worth mentioning that nodes are not limited to immediate neighbors (as denoted by p^+ for one or more property paths). Subsequently, the fine-tuned encoder and decoder are applied to unlabeled nodes V_U. Each candidate pair is scored and added to the set of alignment pairs $\mathcal{G}_{align}$ satisfying the threshold λ.

2.3 Graph Augmentation

In this section we provide details related to **Completion** module, including generating samples from the aligned PHKG with respect to the upper-level pathways. The generated samples

following the upper level ontology concepts and constraints produce the final augmented graph. In particular we focus on generating samples missing from the original PHKG along paths pertaining to clinical observations (lab events), findings (diagnosis), and substances (prescriptions). The samples encode domain knowledge and satisfy ontological constraints with respect to the DSRO as described in the previous section. The generated samples form an **Augmented Graph**, which may be used to complete the information from original input tables, as shown in Figure 1.

More specifically, the augmented graph is generated for each admission following the pathways that connect observations taken during lab events, leading to outcome based diagnoses and prescriptions. These core concepts form the sampling meta-paths, informing the learning process used in generating embeddings by the GCN encoder. Given the range information for each relation along a meta-path edge, the GCN decoder can be used to predict target node types. The predicted types capture the information codified from the DSRO and can in turn be translated into original table values.

The steps for generating a set of N node types following L meta-paths denoted by $\{p_1, \cdots, p_L\}$ are summarized in Algorithm 3. The unlabeled entities $\{o_1, \cdots, o_N\}$ correspond to missing values in the original table, initialized randomly to begin with. Following the GCN training algorithm, for each relation p on the meta-pathway we sample the p-neighborhoods including the unlabeled entities. The encoder and decoder are fine-tuned on these neighborhoods by decreasing the distance between the representations of path-wise neighbors. Once the embedding representations are updated, for each unlabeled node a score s is computed with respect to the relation type p it appears in (as range). The node type is added to the augmented graph $\mathcal{G}_{aug}$ if it satisfies a threshold value λ.

Algorithm 3 Graph Augmentation with Meta-Path Sampling

Input: $\mathcal{G}_{\text{align}}$ for patient u, L meta-paths $\{p_1, \ldots, p_L\}$, set of N blank nodes for augmentation $\{o_1, \ldots, o_N\}$
Output: Augmented graph $\mathcal{G}_{\text{aug}}$

 # Meta-path sampling
Initialize: empty graphs $\{\mathcal{G}_1, \ldots, \mathcal{G}_L\}$
for each predicate p_l in $\{p_1, \ldots, p_L\}$ **do**
  $\mathcal{G}_l \leftarrow \mathcal{G}_{l-1} \cup \{(s, p_l^+, o) \in \mathcal{G}_{\text{align}}\}$
    $\cup \{(s, p_l^+, o) \mid o \in \{o_1, \ldots, o_N\}\}$
  # Update encoder & decoder
  $\text{ENC}, \text{DEC} \leftarrow \text{DEC}(\text{ENC}(\mathcal{G}_l), \text{ENC}(\mathcal{G}_l))$
end for

 # Augmentation step
for v_j in $\{o_1, \ldots, o_N\}$ **do**
  for each predicate p in $\{p_1, \ldots, p_L\}$ **do**
    for v_i in $\{(v_i, p, v_j) \in \mathcal{G}_{\text{align}}\}$ **do**
      $s \leftarrow \text{DEC}(\langle v_i, v_j \rangle)$ ▷ score
      if $s > \lambda$ **then** ▷ threshold
        $\mathcal{G}_{\text{aug}} \leftarrow \mathcal{G}_{\text{aug}} \cup \{(v_i, p, v_j)\}$
      end if
    end for
  end for
end for
return $\mathcal{G}_{\text{aug}}$

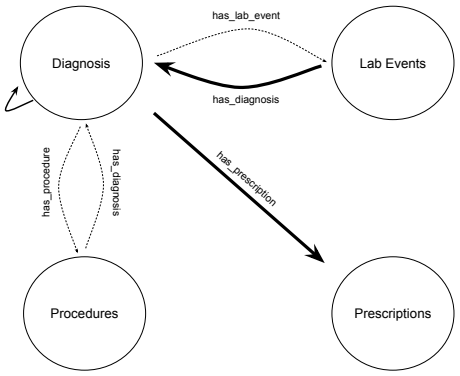


Figure 4: Clinical Upper-level Concepts and Meta-pathways. The highlighted edges indicate the causal paths that inform the use case in our work.

3 Experimentation

In this section, extensive experiments are conducted and reported for evaluating the proposed framework towards aligning and completing tabular EHR records. We report on dataset pre-processing steps, semantic annotation accuracies, ontology alignment results, and predictive performance for lab event detection, diagnosis prediction, and drug recommendation through precision (P), recall (R), and f1 scores (F).

3.1 Datasets

We use the MIMIC repository containing tabular data for patients to ultimately generate triples for

training and evaluation purposes. In this work, we limit the scope to records from patients that are hospitalized for Cardiovascular Disease (CVD). The relevant data are separated by admissions encoded by ICD-9 code range 410-430, such as 428.22 (Chronic systolic heart failure), 428.23 (Acute on chronic systolic heart failure), 428.32 (Chronic diastolic heart failure), 428.33 (Acute on chronic diastolic heart failure), 428.42 (Chronic combined systolic and diastolic heart failure), and 428.43 (Acute on chronic combined systolic and diastolic heart failure). These codes categorize various forms and severities of heart failure based on the systolic and diastolic dysfunction of the heart. In ICD-10, these codes are largely replaced by categories under I50 (Heart Failure). To generate this subset, we identify and store the admissions for those patients that have at least one of the above ICD codes associated with them and exclude items outside the above scope for our final set of patients.

The tabular data used in this work are selected and organized around four core themes, namely Diagnosis, Procedures, Prescriptions, and Lab Events. Although there are cases where extra information such as transfers, provider source, and notes exist, for the purposes of tabular processing related to our use case we organize the data under aforementioned core concepts. These four concepts provide the pathways for most critical care decision making (Mao et al., 2022). In particular, lab events and procedures typically inform diagnosis, while diagnosis decisions inform prescriptions, causally speaking, as shown in Figure 4.

In order to transform tabular data to knowledge graph representation, SPHN[4] is used as a schema that defines core concepts and predicates for modeling clinical patient records (i.e. EHR). In particular, we focus on 13 core concepts, namely, 'LabTestEvents', 'LabResult', 'Code', 'DrugPrescription', 'Drug', 'Substance', 'Diagnosis', 'BilledProcedure', 'AdministrativeCase', 'SubjectPseudoIdentifier', 'MedicalProcedure', 'BodySite', and 'AdministrativeGender'. We also consider an additional concept named 'Patient' in order to model the individual patients. As for predicates, we model a total of 7, namely

[4] https://www.biomedit.ch/rdf/sphn-schema/sphn

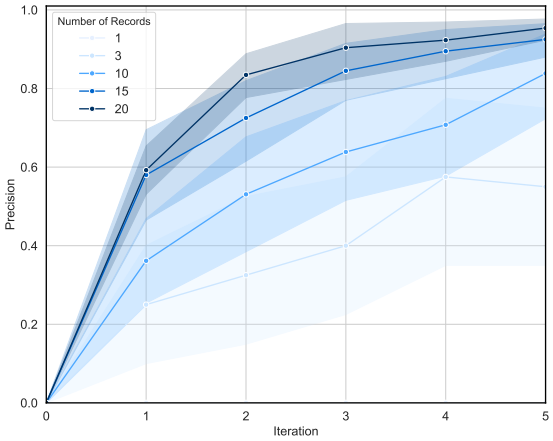


Figure 5: The effect of record quantities available to generate the personal health knowledge graph (PHKG) using LLM Semantic Annotation.

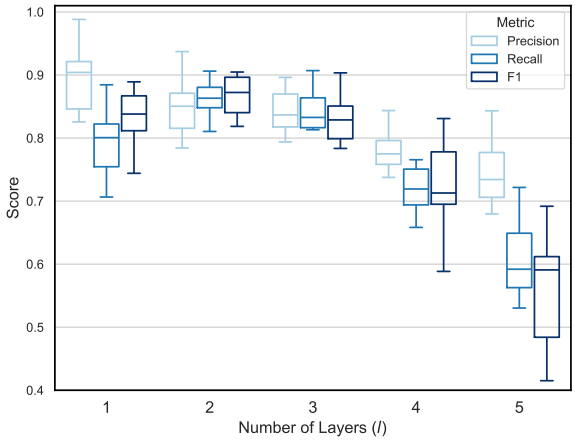


Figure 6: Ontological Alignment scores in terms of number of layers in the Encoder/Decoders.

'hasCode', 'hasLabTest', 'hasAdministrativeCase', 'hasSubjectPseudoIdentifier', 'hasDrug', 'hasActiveIngredient', and 'hasAdministrationRoute' to capture the relations between the entities. Additionally, we include 'is a' relation to indicate the type assertions, 'rdfs:subClassOf' to indicate membership, and 'owl:sameAs' to indicate equivalent codes.

3.2 Results

In the first set of experiments, we demonstrate the effectiveness of the proposed semantic annotation step (i.e. Algorithm 1) for predicting core concepts in the BGP. We run the experiment for upto 5 iterations and measure the predictive precision with 1, 3, 10, 15, and 20 records per core concept, as shown in Figure 5. We observe that with 10 or higher number of records and after 5 iterations, the algorithm achieves satisfactory results. Once the entities are annotated, the PHKG is generated in

triple format.

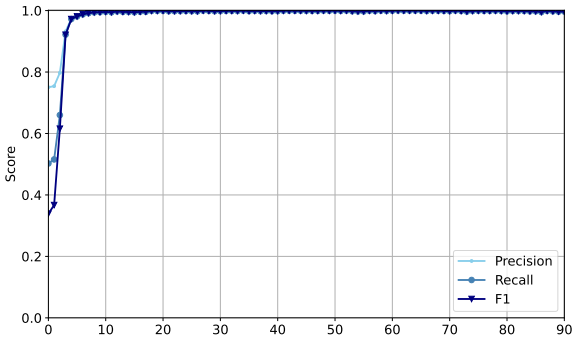
The PHKG embeddings are learned with l convolution operators, each followed by a ReLu and Dropout ($p = 0.2$) layer using the PyGeometric library[5]. The hyperparameters are set by default to `batch_size=1024`, `learning_rate=0.005`, `dropout=0.2`, and `regularization=1e-2`. In our experiments we create a separate train and test split for each task at a random 80-to-20 ratio and train a new model each time.

In order to find the effective model depth for alignment and completion tasks, we run algorithm 2 with different number of layers $l = \{1, 2, 3, 4, 5\}$ of the encoder and measure the predictive precision, recall, and f-1 score of the outcomes at threshold level $\lambda = 0.5$. We observe as shown in Figure 6 that the models achieve the best results up to and including three layers, past which the performance begins to degrade. In the following we set this hyper-parameter as $l = 2$.

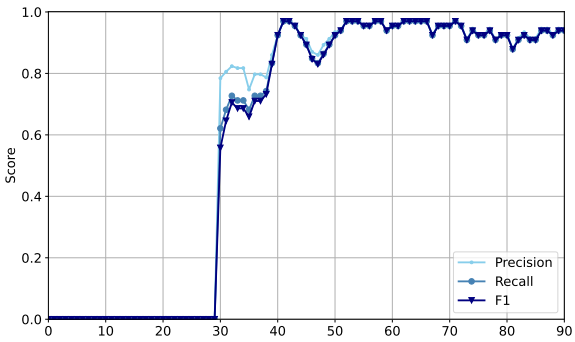
The PHKG contains entities from one or multiple coding systems - ICD for Diagnosis and Procedures, LOINC for Lab and Observation results, and NDC for Drugs and Substances. On the other hand, SNOMED CT enables an encompassing representation of clinical concepts including diagnoses, procedures, observations and substances. Aligning ICD, LOINC, and NDC vocabularies to SNOMED CT allows the encoding of patient data with contextualized representations under one coding scheme, deemed crucial for predictive tasks which we evaluate next.

In Figure 7, we demonstrate the results of meta-path informed generative sampling in terms of precision, recall, and f1-score according to Algorithm 3. The progression of pathways follows 'has lab code' for LOINC code prediction, 'has diagnosis code' for ICD code prediction, and 'has drug code' for NDC drug prediction. For each meta-path, the encoder and decoder are updated for 30 iterations (i.e. a total of 90 iterations). It can be observed with introduction of each new pathway, that the scores exhibit a step function behavior before converging within a window of 20 iterations. All in all, f1-scores of 0.984 , 0.862 , and 0.997 are achieved in this experiment for lab

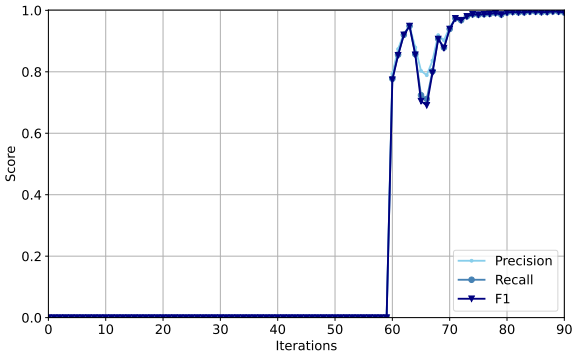
[5]https://pyg.org/



(a) Lab Events



(b) Diagnosis



(c) Prescriptions

Figure 7: Performance scores on test data using the meta-path sample generation of Algorithm 3, augmenting a personal health knowledge graph including 100 random admissions and following lab event (p_1), diagnosis (p_2), and prescription (p_3) pathways.

event, diagnosis, and prescription code imputation.

We experiment further and report results for various down-stream prediction tasks using our graph augmentation framework in Table 1. We provide performance details in terms of three tasks, namely lab event detection, diagnosis prediction, and drug recommendation. Each task is defined as predicting the corresponding code given the embedded and aligned context from a particular admission of a test patient. We experiment with both the

Dataset Size	Drug Recommendation			Lab Event Detection			Diagnosis Prediction		
	P	R	F	P	R	F	P	R	F
MIMIC III									
DS100	0.99 ± 0.002	0.99 ± 0.002	0.99 ± 0.002	0.92 ± 0.009	0.91 ± 0.012	0.91 ± 0.013	0.87 ± 0.018	0.85 ± 0.022	0.85 ± 0.023
DS200	0.99 ± 0.002	0.99 ± 0.002	0.99 ± 0.002	0.87 ± 0.018	0.83 ± 0.031	0.82 ± 0.034	0.96 ± 0.009	0.96 ± 0.009	0.96 ± 0.009
DS300	0.99 ± 0.002	0.98 ± 0.002	0.98 ± 0.002	0.83 ± 0.010	0.73 ± 0.024	0.71 ± 0.029	0.96 ± 0.013	0.96 ± 0.014	0.96 ± 0.014
DS400	0.98 ± 0.002	0.98 ± 0.002	0.98 ± 0.002	0.84 ± 0.016	0.77 ± 0.033	0.76 ± 0.039	0.94 ± 0.021	0.94 ± 0.021	0.94 ± 0.021
DS500	0.98 ± 0.001	0.98 ± 0.001	0.98 ± 0.001	0.84 ± 0.012	0.78 ± 0.025	0.76 ± 0.029	0.97 ± 0.007	0.97 ± 0.008	0.97 ± 0.008
Average	0.99 ± 0.005	0.98 ± 0.005	0.98 ± 0.005	0.86 ± 0.034	0.80 ± 0.064	0.79 ± 0.074	0.94 ± 0.039	0.94 ± 0.048	0.94 ± 0.048
MIMIC IV									
DS100	1.00 ± 0.002	1.00 ± 0.002	1.00 ± 0.002	0.95 ± 0.010	0.94 ± 0.013	0.94 ± 0.013	0.93 ± 0.011	0.92 ± 0.014	0.91 ± 0.014
DS200	0.99 ± 0.002	0.99 ± 0.002	0.99 ± 0.002	0.89 ± 0.021	0.85 ± 0.035	0.85 ± 0.038	0.98 ± 0.014	0.98 ± 0.015	0.98 ± 0.015
DS300	0.98 ± 0.004	0.98 ± 0.004	0.98 ± 0.004	0.89 ± 0.020	0.85 ± 0.033	0.85 ± 0.036	0.96 ± 0.014	0.96 ± 0.015	0.96 ± 0.015
DS400	0.98 ± 0.003	0.98 ± 0.003	0.98 ± 0.003	0.83 ± 0.020	0.74 ± 0.048	0.72 ± 0.061	0.94 ± 0.021	0.94 ± 0.022	0.94 ± 0.022
DS500	0.98 ± 0.002	0.98 ± 0.002	0.98 ± 0.002	0.87 ± 0.015	0.82 ± 0.027	0.81 ± 0.030	0.96 ± 0.010	0.96 ± 0.011	0.96 ± 0.011
Average	0.99 ± 0.008	0.99 ± 0.008	0.99 ± 0.008	0.89 ± 0.042	0.84 ± 0.070	0.83 ± 0.082	0.95 ± 0.019	0.95 ± 0.021	0.95 ± 0.025

Table 1: Performance evaluation of the proposed meta-path sampling generation algorithm for predictive tasks, i.e. Drug Recommendation, Lab Event Detection, and Diagnosis Prediction. Different sizes of datasets are used, including 100 to 500 admissions in each case (DS100-DS500) from both MIMIC III and MIMIC IV. Mean and standard deviation over 10 separate runs are reported, in terms of precision (P), recall (R), and f1 score (F).

third and forth version of the MIMIC repository (MIMIC III and MIMIC IV) and run the experiment with different input sizes. In particular, we generate graphs with randomly sampled data from 100, 200, 300, 400, and 500 distinct admissions (i.e. DS100-DS500). It can be observed that the models consistently achieve high performance in precision, recall, and f1 score for each prediction task and across different graph sizes.

4 Conclusions

In this work, a framework is proposed that transposes the electronic health records from real-world patients in tabular format with graphical representation using generative sampling. The representations are aligned with a domain specific ontology to further disambiguate and contextualize. A graph neural network that supports multi-relational entities is trained and meta-path sampling is applied to generate missing information according to upper-level ontological information. The generation process applied to tabular inputs related to cardiovascular disease, achieve precision, recall, and f1 scores in the ideal range for clinical data augmentation and decision making.

References

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. 2020. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 606–613.

Ioannis Dasoulas, Duo Yang, Xuemin Duan, and Anastasia Dimou. 2023. Torchictab: semantic table annotation with wikidata and language models. In *CEUR Workshop Proceedings*, pages 21–37. CEUR Workshop Proceedings.

Junheng Hao, Chuan Lei, Vasilis Efthymiou, Abdul Quamar, Fatma Özcan, Yizhou Sun, and Wei Wang. 2021. Medto: Medical data to ontology matching using hybrid graph neural networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 2946–2954, New York, NY, USA. Association for Computing Machinery.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Chengsheng Mao, Liang Yao, and Yuan Luo. 2022. Medgcn: Medication recommendation and lab test imputation via graph convolutional networks. *Journal of Biomedical Informatics*, 127:104000.

Shervin Mehryar. 2025. A resolution-alignment-completeness system for data imputation over tabular clinical records. In *ELLIS workshop on Representation Learning and Generative Models for Structured Data*. ELLIS workshop on Representation Learning and Generative Models for Structured Data.

Joshua Robinson, Rishabh Ranjan, Weihua Hu, Kexin Huang, Jiaqi Han, Alejandro Dobles, Matthias Fey, Jan Eric Lenssen, Yiwen Yuan, Zecheng Zhang, and 1 others. Relational deep learning: Graph representation learning on relational databases. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

Fei Teng, Wei Yang, Li Chen, LuFei Huang, and Qiang Xu. 2020. Explainable prediction of medical codes with knowledge graphs. *Frontiers in bioengineering and biotechnology*, 8:867.

Xiang Yue, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yungui Huang, Simon M Lin, Wen Zhang, Ping Zhang, and Huan Sun. 2020. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics*, 36(4):1241–1251.