# Multiview Incomplete Knowledge Graph Integration with application to cross-institutional EHR data harmonization

Link: <https://pubmed.ncbi.nlm.nih.gov/35872266/>

**Motivation**

What is the problem being solved?

To integrate information from multiple sources with partially overlapping EHR concept codes to enable translations between healthcare systems

Why is it important?

* There is a lack of interoperability across healthcare systems.
  + differential coding practices between healthcare systems
  + financial incentives between healthcare systems
  + same clinical feature might be represented by distinct codes at different healthcare systems
* Mappings from local codes to standardized ontologies tend to be incomplete, ambiguous, inaccurate
  + Ontologies are constantly updated over time
  + Comes with complicated relationships such as one-to-many and many-to-many mapping
* Time-consuming, resource-intensive
  + subjective bias and human errors.
  + rely on domain knowledge

What previous work exists?

* Knowledge graph (KG)
  + Translate components of a KG, which are essentially entities and their relations, to lower dimensional embedding vectors
* Neural network based methods
  + Train a shallow neural network to predict codes or context of codes
  + Cosine similarity -> find nearest neighbors of a code -> Dictionary that maps codes between two institutions
  + skip-gram algorithm
  + matrix factorization (SPPMI-SVD)

Why is the previous work insufficient to solve the problem?

* KG
  + require input data to be structured as triples of the form
  + embeddings are typically trained from well structured knowledge databases
  + free form unlabeled EHR data cannot be easily turned into such triplets
* Neural network based
  + use patient-level EHR data
  + privacy concerns
* Matrix factorization-based embedding algorithms
  + does not incorporate information from code textual descriptions

**Approach**

SAPBERT (textual information)

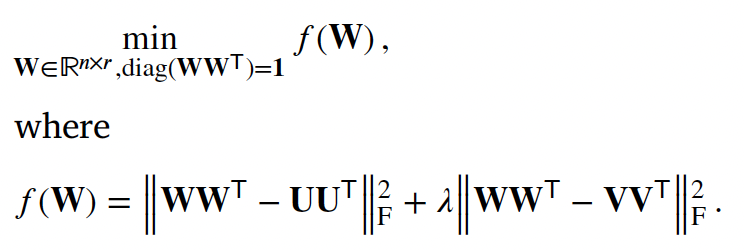
* generate embeddings by extracting semantic information from code descriptions
* distinct embeddings for codes that are unique to two institutions
* same embeddings for codes that are shared by the institutions
* normalized unit-length embeddings

SPPMI-SVD (co-occurrence)

* obtain two initial sets of embeddings based (one each institution)
* synchronize the two sets of embeddings by generating an alternative representation of all codes
  + based on its cosine similarity with the overlapping codes
  + compute the cosine-similarity matrices
  + SynC embeddings: combine to form final set of synchronized cosine-similarity based embeddings

MIKGI

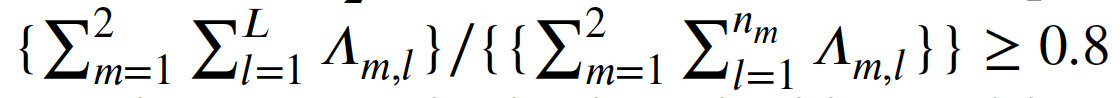
* integrate information from the code descriptions and co-occurrence patterns
* by solving a constrained minimization problem using PGD



* first component: textual info; second component: co-occurrence info

Tuning parameters

* SAPBERT dimension
* SPPMI-SVD dimensions (for each initial institution embeddings): assume to be the same
  + Smallest L s.t.



* MIKGI regularization lambda and dimension
  + maximize the accuracy of the trained embeddings 𝐖 in detecting known relationship pairs (AUC)

**Validation**

Datasets

* Veterans Affairs (VA)
* Healthcare System and the Mass General Brigham (MGB)

Detecting similar or related concepts

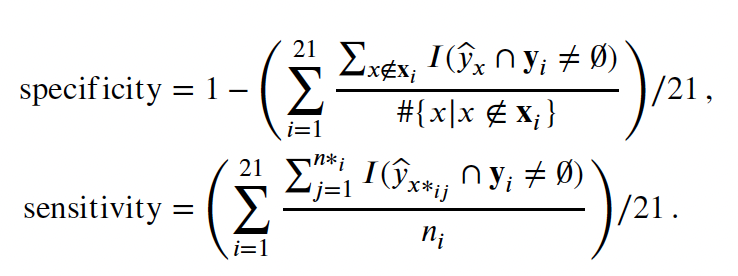
* Similar pairs
  + pairs of codes that represent highly similar clinical concepts according to existing ontologies
  + leverage ontologies’s hierarchy to define similar pairs + 916 similar lab–lab pairs manually curated by domain experts
* Related pairs
  + pairs of codes that have more complex relation like ‘may cause’ or ‘may treat’.
  + curate known relationship pairs from online knowledge sources
* Evaluation
  + cosine similarities of the embedding vectors of known pairs and those of randomly selected pairs
  + AUC of the cosine similarities in distinguishing known pairs from random pairs

Code mapping across two institutions

* local VA medication codes to RxNorm
  + compare against manually curated labels
  + Evaluation: based on the top 𝑘 accuracy (acc@k)
* local VA lab codes to MGB LOINC and/or LP codes
  + train supervised algorithms based on the standard orthogonal transform algorithm
  + compare to a maximum entropy (Max) based supervised algorithm
    - use normalized local lab descriptions as input features by generating a set of tokens using lexical rules and the corresponding LOINC codes of each local codes as labels
  + Evaluation: acc@𝑘 (excluding the 500 training pairs)

Case study of selecting VA local lab codes for a COVID-19 EHR study

* manually curated all VA lab codes and MGB LOINC and LP codes that should be mapped to 21 laboratory tests
* Evaluation:

****

* x = VA lab code; x\_\*i = all such lab codes except curated lab concepts for the 𝑖th test
* y =LOINC or LP codes; y\_hat = top 𝑘 LOINC or LP codes that 𝑥 is mapped to

Identifying features important for COVID-19

* examine whether MIGKI can effectively identify relevant signs/symptoms, medications, laboratory tests to study COVID-19 as a novel disease

**Results**

Detecting similar or related concepts

* MIKGI is the only algorithm that attained the highest or near the highest performance for both
  + similarity (AUC = 0.918)
  + relatedness (AUC = 0.809)

Cross-institutional code mapping

* MIKGI attained a higher accuracy for lab code mapping and medication to RxNorm mapping
* Using the supervised training with 500 labels, the lab code mapping accuracy improved substantially for MIKGI

| Mapping | acc@1 | acc@5 | acc@10 | acc@20 |
| --- | --- | --- | --- | --- |
| VA LAB Code → LOINC/LP | 0.591 | 0.758 | 0.818 | 0.871 |
| VA Lab Code → LOINC/LP  supervised learning with 500 labels | 0.777 | 0.879 | 0.904 | 0.929 |
| VA Medication Code → RxNorm | 0.910 | 0.975 | 0.995 | 0.995 |

Case study of selecting VA local lab codes for a COVID EHR study

* MIKGI achieved the highest sensitivity among all methods while maintaining a high specificity

Performance in identifying COVID-19 related codes

* Identified 167 original EHR codes and 25 LP codes as potentially relevant for COVID-19
* able to identify key symptoms, medications, procedures
* consistent with recent literature on the diagnosis and management of COVID-19

**Contributions**

* Generates high quality embeddings
  + simultaneously represent EHR codes from multiple institutions
  + integrating information from code descriptions and co-occurrence patterns
* Overcomes data privacy challenges
  + enables cross institutional code mapping for data harmonization
  + MIKGI-based code embeddings can be used to directly project patient features to an embedding space
  + can facilitate multi-institutional integrative analysis and improve cross institutional transportability
* Attains the most robust performance across multiple downstream tasks compared to the commonly used embedding methods
  + PubmedBERT, CODER, OT, and PT
    - either do not incorporate code description or only use short phrases with insufficient information
  + SAPBERT, BioBERT:
    - better than the other 4
    - MIKGI still attains more robust performance than them by further leveraging code co-occurrence patterns in EHR data, which essentially encode the meaning of codes in clinical practice based on their relationship with other codes.
* Computationally efficient
* Can be updated over time
  + only relies on the simple co-occurrence matrix of code pairs from the EHR along with code descriptions
* Serve as feature selection tools or directly representing selected features
* Can be easily adapted to include >2 EHR systems/institutions
  + multi-task matrix completion problem
  + gradient descent algorithm

**Limitations**

Unsupervised VA lab code mapping only attained acc@1 of 59%

* the challenge of lab code descriptions having varying degrees of ambiguity with the use of acronyms
* not fully utilizing the EHR information on the laboratory tests such as their findings (e.g., high vs normal)
* certain laboratory tests are always ordered together as part of a panel, which results in identical EHR embeddings for their associated EHR codes

What are the next steps for future research?

* Further improve the code mapping accuracy via supervised methods
* Other supervised or semi-supervised machine learning methods can be used
* Further enrich the knowledge network by predicting specific associations between a pair of EHR concepts such as ‘‘may treat’’ or ‘‘may cause’’