MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

Graph Representation Learning-based Fixed-length Clinical Feature Vector Generation from Heterogeneous Medical Records

@tmhsskl

Tomohisa Seki

Research Associate

Department of Healthcare Information Management, The University of Tokyo Hospital

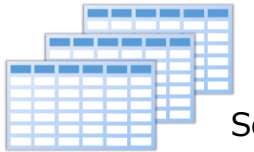
AIDH IMIA

Backgrounds

- The process of retrieving patients from a medical information database, guided by specific medical concepts, can be likened to a manual feature engineering procedure.
- The tasks of data extraction and preprocessing, contingent upon medical and medical informatics expertise, necessitate advanced knowledge and specialized skillsets.
- Utilizing machine learning for automated feature extraction offers significant potential, particularly if this could be adapted for feature extraction from medical data, harmonious with medical concepts.

Backgrounds

Current medical data processing process

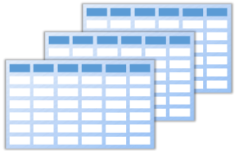


Selection and consolidation of information based on expert knowledge

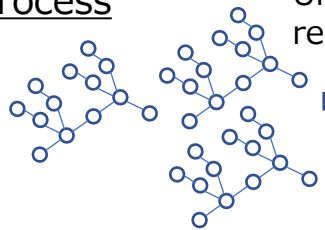
missing value processing and imputation

Table data

Proposed medical data processing process



Conversion to graph data



Unsupervised graph representation learning

Vector data containing clinical characteristics

```
[0.53288564 0.03550337 0.95084103 0.27537568 0.01618467]
[0.07644401 0.5801949  0.13778552 0.04172479 0.37962647]
[0.62276119 0.1118398  0.31332783 0.41776067 0.1594601 ]
[0.5060739  0.08521409 0.0465185  0.51983537 0.24789697]
                          ⋮
[0.09537381 0.42391259 0.91514074 0.72419245 0.81561592]
[0.87191399 0.97262774 0.76324657 0.60840792 0.36542588]
```

Should the medical information be transformed into graph structures, enabling graph representation learning as opposed to conventional tabular data formation, it could potentially bypass the need for extraction processes heavily dependent on expert knowledge.

Objective

- Our aim is to investigate the potential of unsupervised Graph Representation Learning in extracting features from electronic medical records accumulated during hospitalization, and to determine whether it can obtain a fixed-length embedded representation that effectively retains essential clinical information.

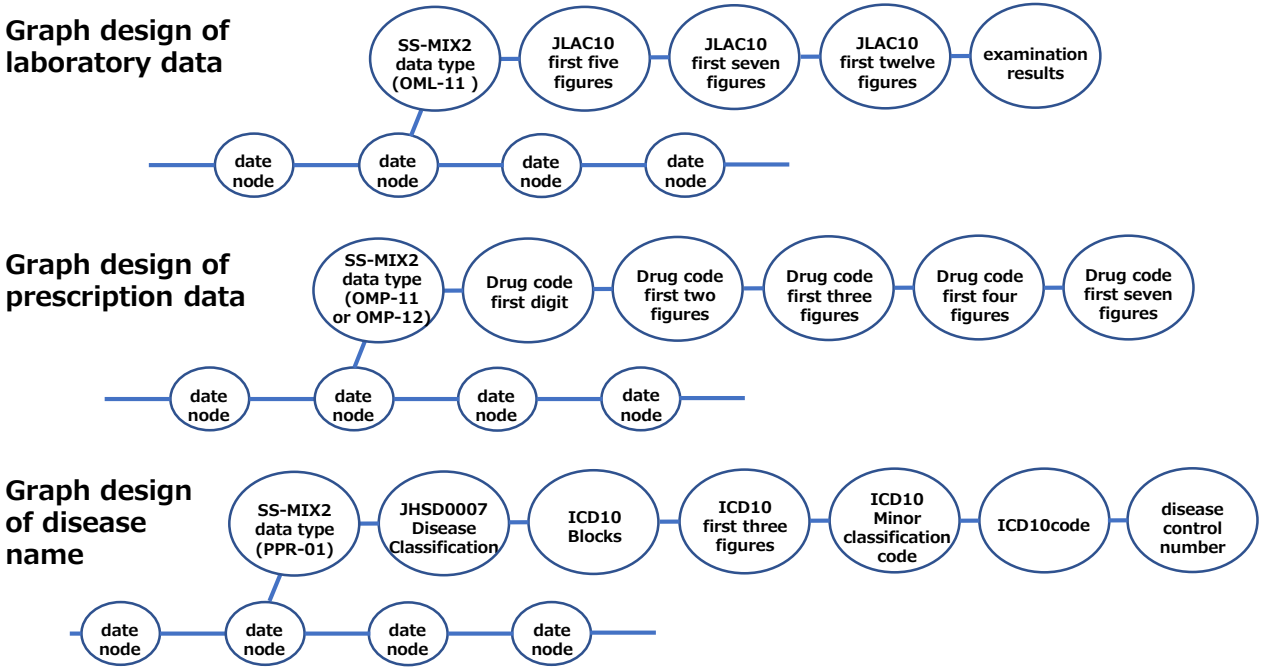
Methods: Data preparation

- This study utilized anonymized SS-MIX2 standardized storage, a Japanese standard for the preservation of HL7 Ver.2.5 text files, from the University of Tokyo Hospital.
- We utilized laboratory tests, prescriptions, and disease information recorded during patients' hospital stays.
- For training, we harnessed data of 31,679 patients and 52,667 instances of hospitalization from the years 2015 and 2016.
- For validation, we employed data from 15,417 patients and 21,763 instances of hospitalization recorded in 2017.

Methods: Conversion to graphs

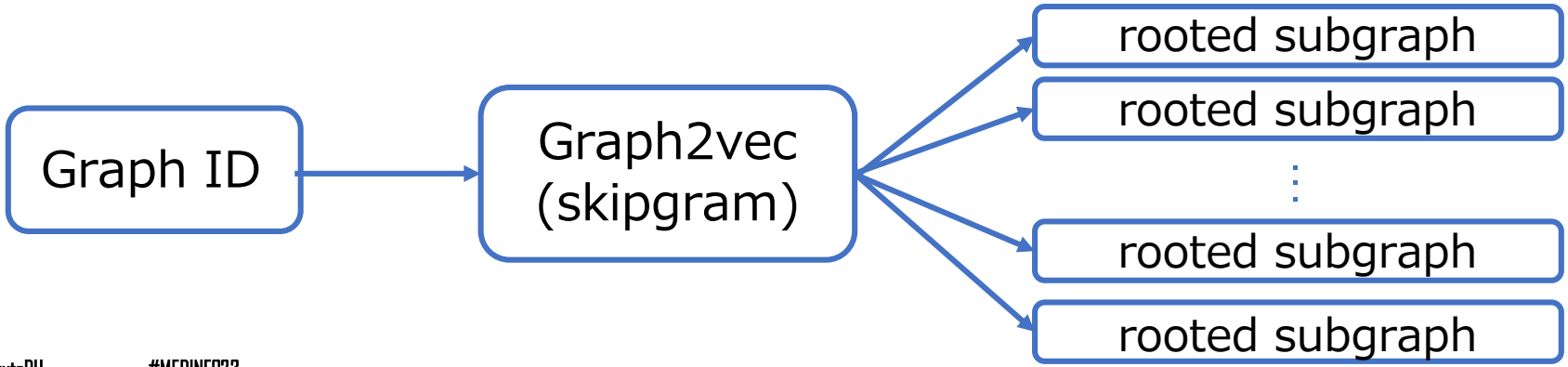
- We devised a coding scheme that allows for partitioning and transformation into graph structures, with the aim of expressing the concurrence of higher-level concepts as node label agreements.

- This method greatly simplifies the process of deriving similarity measures.



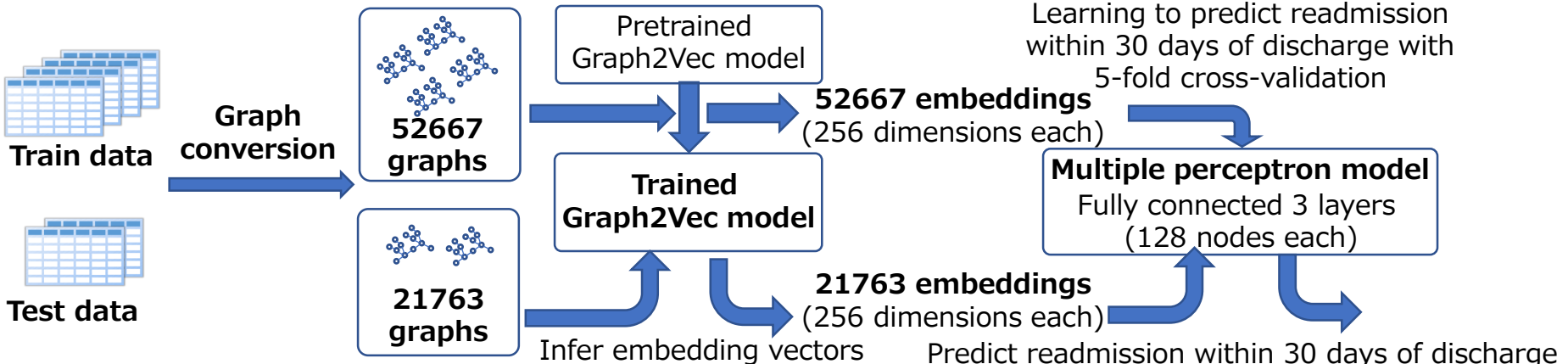
Methods: Graph embedding

- We employed Graph2Vec for our graph representation learning approach.
- Graph2Vec functions by leveraging skipgrams, as utilized in Doc2Vec, to learn a distributed representation of the entire graph.



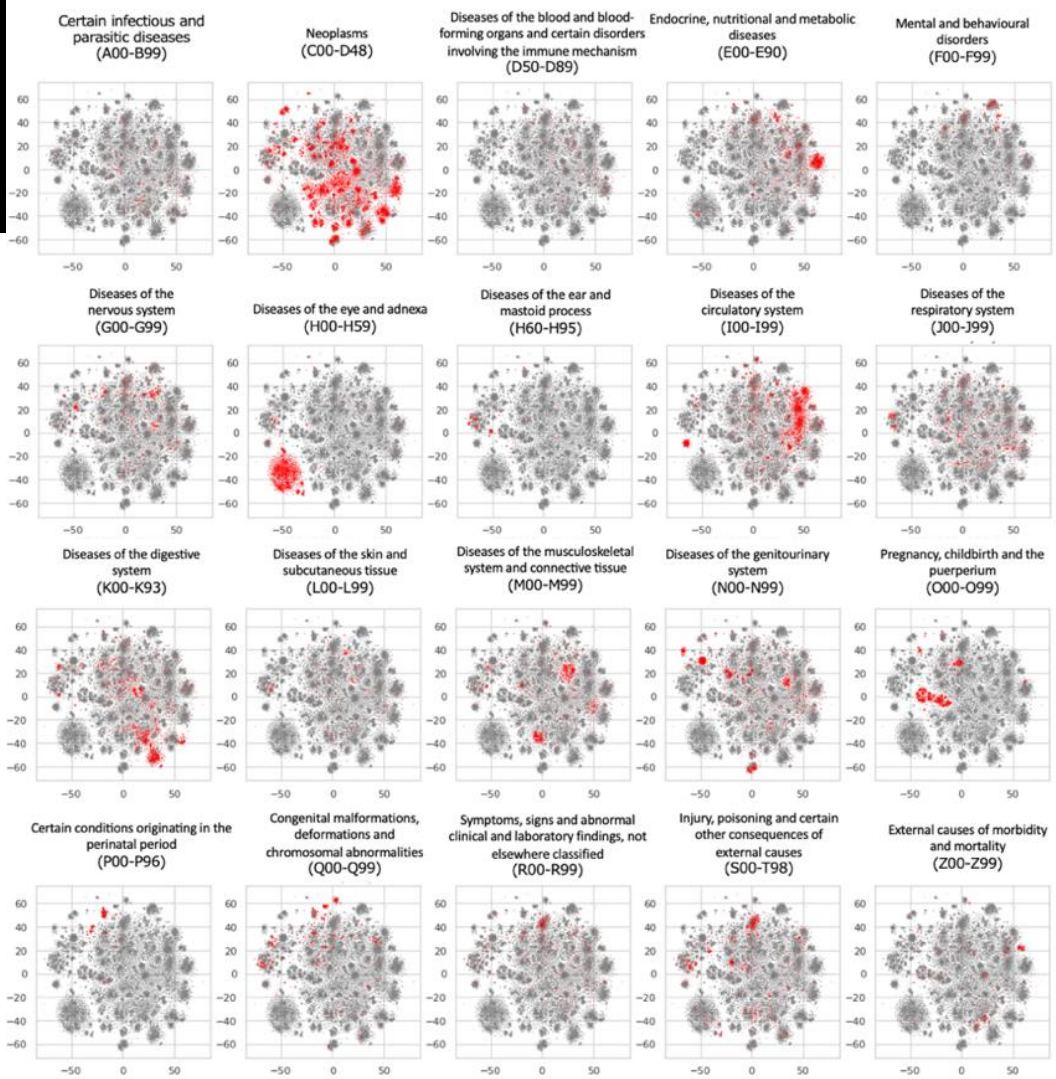
Methods: Evaluation of embedding

- We performed a T-SNE plot analysis, color-coded according to the 20 ICD10 code block classifications. Furthermore, we utilized multilayer perceptron models to predict readmissions within 30 days post-discharge.



Results: T-SNE plot

- The two-dimensional visualization via T-SNE revealed that the embedded representations were distributed unevenly, correlating with the block categories of ICD10 codes designated as the primary diagnoses for hospital admissions.



Results: Prediction of readmission

- Our findings indicate that the prediction performance for readmissions within 30 days following discharge tends to improve as the volume of information within the graph representation escalates.

Included information				Validation results on test data			
Date	Drug	Lab	Diagnosis	AUC (SD)	Precision (SD)	Recall (SD)	F1 score (SD)
+	-	-	-	0.500 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
+	+	-	-	0.570 (0.009)	0.526 (0.021)	0.165 (0.022)	0.250 (0.024)
+	-	+	-	0.587 (0.004)	0.205 (0.010)	0.484 (0.038)	0.288 (0.004)
+	+	+	-	0.609 (0.017)	0.445 (0.031)	0.278 (0.055)	0.337 (0.027)
+	+	+	+	0.627 (0.021)	0.410 (0.027)	0.336 (0.063)	0.363 (0.036)

Discussion

- Unsupervised graph representation learning has effectively produced fixed-length vectors that encapsulate meaningful clinical information derived from irregular medical data.
- To further enhance the utility of these vectors, several considerations remain, including the integration of additional medical data, modifications to the graph structure, and exploration of time-series feature extraction methods.

Conclusion

- Feature extraction from electronic medical records using unsupervised graph representation learning can procure a fixed-length embedded representation that retains critical clinical information.
- Our approach shows potential for automating the extraction of patient characteristics from electronic medical records, and could serve as a valuable tool in managing electronic medical records.

Co-authors

- Yoshimasa Kawazoe [a,b]
- Kazuhiko Ohe [a,c]

a Department of Healthcare Information Management, The University of Tokyo Hospital

b Artificial Intelligence in Healthcare, Graduate School of Medicine, The University of Tokyo

c Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo