Comparative Effectiveness of Knowledge Graphs- and EHR Data-Based Medical Concept Embedding for Phenotyping

Junghwan Lee‡, Cong Liu‡*, Jae Hyun Kim, Alex Butler, Ning Shang, Chao Pang, Karthik Natarajan, Patrick Ryan, Casey Ta$, Chunhua Weng$
Department of Biomedical Informatics, Columbia University, New York, NY 10032

‡Equal contribution first authors; $Equal-contribution senior authors

*Corresponding Author:
Cong Liu, PhD
Department of Biomedical Informatics
Columbia University
622 W 168 Street, PH-20, Room 407, New York, NY 10032
Email: cl3720@cumc.columbia.edu

Abstract word count: 250
Word count: 3,951

KEYWORDS
Embedding,
Representation Learning,
Phenotyping,
Knowledge Graph,
Electronic Health Records

ABSTRACT

Objective: Concept identification is a major bottleneck in phenotyping. Properly learned medical concept embeddings (MCEs) have semantic meaning of the medical concepts, thus useful for feature engineering in phenotyping tasks. The objective of this study is to compare the effectiveness of MCEs learned by using knowledge graphs and EHR data for facilitating high-throughput phenotyping.

Materials and Methods: We investigated four MCEs learned from different data sources and methods. Knowledge-graphs were obtained from the Observational Medical Outcomes Partnership (OMOP) common data model. Medical concept co-occurrence statistics were obtained from Columbia University Irving Medical Center's (CUIMC) OMOP database. Two embedding methods, node2vec and GloVe, were used to learn embeddings for medical concepts. We used phenotypes with their corresponding concepts generated and validated by the Electronic Medical Records and Genomics (eMERGE) network to evaluate the performance of learned MCEs in identifying phenotype-relevant concepts.

Results: *Precision@k%* and *Recall@k%* in identifying phenotype-relevant concepts based on a single concept and multiple seed concepts were used to evaluate MCEs. *Recall@500%* and *Precision@500%* based on a single seed concept of MCE learned using the enriched knowledge graph were 0.64 and 0.13, compared to *Recall@500%* and *Precision@500%* of MCE learned using the hierarchical knowledge graph (0.61 and 0.12), 5-year windowed EHR (0.51 and 0.10), and visit-windowed EHR (0.46 and 0.09).

Conclusion: Medical concept embedding enables scalable identification of phenotype-relevant medical concepts, thereby facilitating high-throughput phenotyping. Knowledge graphs

constructed by hierarchical relationships among medical concepts learn more effective MCEs, highlighting the need of more sophisticated use of big data to leverage MCEs for phenotyping.

INTRODUCTION

Phenotyping is a task of identifying a patient cohort underlying specific clinical characteristics[1]. With the widespread adoption of electronic health records (EHR), phenotyping is one of the most fundamental research challenges encountered when using the EHR data for clinical research[2]. As learned from the Electronic Medical Records and Genomics (eMERGE) network, the process of developing and validating a phenotype requires a large amount of manual effort and time, typically up to 6-10 months[3]. An essential but often labor-intensive step is to identify phenotype-relevant medical concepts (i.e. feature engineering) such as relevant diagnoses, laboratory tests, medications, and procedures. A typical phenotype can contain thousands to tens of thousands of medical concepts. For example, the Type 2 Diabetes Mellitus (T2DM) phenotype developed by the eMERGE network contains about 12,000 relevant medical concepts[4,5]. Feature engineering for rule-based phenotyping largely relies on domain experts and can be error-prone and not generalizable or portable. Data-driven high-throughput phenotyping methods have been proposed to extract relevant features from external knowledge sources[6-10], with neural embedding, which transforms the original data into a vector representation and another feature dimension. Properly learned embeddings capture the underlying semantic meaning of the features and hence can be leveraged in phenotyping[11]. There have been numerous efforts to learn efficient medical concept embeddings (MCEs)[12,13] and use them to improve the performance of various tasks such as patient visit prediction[14,15], risk prediction[16], and mortality prediction[17]. This study investigates the comparative effectiveness of using knowledge graphs and EHR data to strengthen MCEs for phenotyping.

Knowledge graphs are one of the widely used resources for learning medical concept embeddings. A knowledge graph contains medical concepts as nodes, which are connected via

various relationships defined according to domain knowledge. Common knowledge graphs include Unified Medical Language System (UMLS), Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT), International Classification of Disease (ICD), and Human Phenotype Ontology (HPO). Graph-based embedding techniques have been leveraged to capture the diversity of connectivity patterns observed in graphs to learn the node embedding for medical concepts[18-20]. For example, Choi et al. proposed Graph-based Attention Model (GRAM), which learns efficient embeddings of medical concepts and applies them for sequential diagnosis prediction[21]. A knowledge graph can be enriched by introducing other kinds of relationships to increase connectivity of nodes in the knowledge graph. Shen et al. learned efficient embeddings for HPO concepts by enriching the HPO knowledge graph leveraging heterogeneous vocabulary resources[22].

EHR are another popular resource to learn medical concept embeddings. A patient visit triggers a number of medical concepts being documented in the EHR. A typical strategy to utilize EHR for learning MCE is to consider each visit or pre-determined time window in a patient EHR as a bag of medical concepts to derive concept co-occurrences. Embedding methods using co-occurrence information, such as GloVe[23] and skip-gram[24], can then be adopted to learn MCEs. For example, Choi et al. developed Med2vec, which leveraged co-occurrence information based on patient visits to learn MCE and showed strong performance in patient's future visit and status prediction[25].

Since MCEs can be used to measure the semantic distance between medical concepts, we hypothesize that properly learned MCEs have the potential to provide a scalable approach to cluster relevant concepts together, thereby mitigating feature engineering efforts and accelerating phenotype development. In this study, we compared how effectively MCEs learned from various

data sources can facilitate high-throughput phenotyping. We trained MCEs based on two different data sources – a knowledge graph and concept co-occurrences obtained from EHRs. GloVe and node2vec, which are widely used feature embedding methods, were implemented to train the MCEs. To explore the performance of different MCEs, we built a gold standard dataset containing 33 phenotypes with their corresponding medical concept lists using phenotypes developed and validated by the eMERGE network[26]. We assessed the learned MCEs on identifying phenotype-relevant medical concepts given seed concepts extracted from a phenotype.

MATERIALS AND METHODS

Data description and processing

All concepts used in this study are based on the Observational Health Data Science and Informatics (OHDSI) Observational Medical Outcomes Partnership common data model (OMOP CDM). OHDSI is a multi-stakeholder, interdisciplinary collaborative that aims to bring out the value of health data through large-scale analytics[27]. The OMOP CDM harmonizes several different medical coding systems, including but not limited to ICD-9-CM, ICD-10-CM, SNOMED-CT, and LOINC, to achieve standardized vocabularies while minimizing information loss, thus provides a unifying data format for various analysis pipelines[28]. We focus on embeddings for condition (i.e. diagnosis) concepts in this study, which play a critical role in phenotyping.

We compared two kinds of knowledge graphs in this study, hierarchical knowledge graph and enriched knowledge graph. The hierarchical knowledge graph was constructed by using “is-a” and “subsume” relationships between condition concepts from the *concept_relationship* table. The enriched knowledge graph expands upon the hierarchical knowledge graph by adding

hierarchical relationships connected to the existing nodes in the hierarchical knowledge graph from the *concept_ancestor* table. Hierarchical relationships are defined such that the child concept has all the attributes of the parent concept with one or more additional attributes (e.g., a relationship between "*Asthma*" and "*Gasping for breath*"). The additional hierarchical relationships increase the connectivity of the knowledge graph and expand the number of concept nodes in the graph. Basic statistics of the knowledge graphs are summarized in **Table 1**.

Table 1. Basic statistics of the knowledge graphs.

	Hierarchical knowledge graph	**Enriched knowledge graph**
# of unique concepts (concept nodes)	306,266	312,089
# of edges (unique relationships)	588,298	671,591
Avg. # of edges per concept node	1.93	2.15

EHR data used to generate concept co-occurrence statistics were obtained from the Columbia University Irving Medical Center (CUIMC) EHR clinical data warehouse containing inpatient and outpatient data starting from 1985. The CUIMC EHR data has been converted to the OMOP CDM and covers more than 36,000 medical concepts from multiple domains (e.g., condition,

drug, and procedure) extracted from more than 5 million patients. To ensure data quality, we used the recent 5-year data from 2013 to 2017 to calculate co-occurrences between concepts[29]. We processed the EHR data into the format of bag-of-medical concepts by applying visit window and 5-year window on each patient's EHR. **Figure 1** depicts how the 5-year window (Fig. 1B) and visit window (Fig. 1C) were applied to patient records. We excluded patients whose EHRs had only a single concept since they do not provide any meaningful co-occurrence information. Basic statistics of the EHR are summarized in **Table 2**. This study received institutional review board approval (AAAD1873) with a waiver for informed consent.

Table 2. Basic statistics of windowed EHR.

	visit-windowed EHR	**5-year-windowed EHR**
# of patients	1,750,889	
# of windowed records	8,865,691	1,370,787
# of unique concepts	17,175	17,288
Avg. (std) # of concepts per windowed record	4.26 (3.31)	12.4 (17.15)

Methods for learning medical concept embedding

GloVe

GloVe was originally developed in the natural language processing domain for learning word representations[23]. GloVe uses the co-occurrence statistics of words to learn the representations of the words by minimizing Eq (1):

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \widetilde{w}_j + b_i + \tilde{b}_j - log X_{ij})^2 \qquad Eq\ (1)$$

where X_{ij} is the number of times word j occurs in the context of word i, V is the size of the vocabulary, and w_i, $\widetilde{w}_j$, b_i, and $\tilde{b}_j$ are word vectors, context word vectors, bias for word vectors, and bias for context word vectors, respectively. Although GloVe was designed to learn word representations, by treating medical concepts as words and patient records as a bag-of-concepts, we can leverage GloVe to learn the representations of medical concepts.

node2vec

node2vec learns representations for nodes in a graph using random walk to optimize the representations[30]. The representations of nodes learned by node2vec have information regarding homophily and structural equivalence of the nodes by adopting two random walk parameters that govern breadth-first search and depth-first search. node2vec learns the representations of the nodes in a graph by maximizing Eq (2):

$$f = \sum_{u \in V} log\ Pr(N_s(u)|f(u)) \qquad Eq(2)$$

where u is a target node, V is the set of all nodes, $N_s(u)$ is the network neighborhood of node u, and f is a mapping function that maps nodes to feature representations.

Medical concept embedding

graphEmb and graphEmb+

graphEmb and graphEmb+ are both embeddings based on a knowledge graph that were trained by implementing node2vec with python 3.5.1 (node2vec package is available at https://github.com/aditya-grover/node2vec). *graphEmb* was trained using the hierarchical knowledge graph, which has 306,266 concepts. *graphEmb+* was trained using the enriched knowledge graph, which has 312,089 concepts. All hyperparameters were set in accordance with the original *node2vec* publication[30], including a single training epoch.

visitEmb and 5yearEmb

visitEmb and 5yearEmb are both concept co-occurrence based embeddings that were trained by implementing GloVe with TensorFlow 2.2.0[31]. *visitEmb* was trained by using co-occurrence statistics between 17,175 concepts using visit windows on patients' records. *5yearEmb* was trained by using co-occurrence statistics of 17,288 concepts using the 5-year window. We limited the maximum co-occurrence value to 1,000 to prevent infrequent co-occurrences from being overshadowed by highly frequent co-occurrences. The MCEs were trained with 50 epochs. Adadelta[32] was used in training with the batch size of 512,000. All other hyperparameters were the same as in the original GloVe publication[23].

Implementation details

All MCEs were trained on a machine equipped with 2 × Intel Xeon Silver 4110 CPUs, 188GB RAM, and 4 × Nvidia GeForce RTX 2080 TI GPUs. The source code is publicly available at https://github.com/WengLab-InformaticsResearch/mcephe. The dimensions of all MCEs were equally set to 128.

Evaluation Strategy

To assess the performance of learned MCEs in identifying relevant medical concept for phenotypes, we used 33 independently validated phenotyping algorithms from the eMERGE network with their corresponding code books. The eMERGE network created the Phenotype Knowledgebase (PheKB) to facilitate phenotyping and sharing of the phenotyping knowledge[26], where phenotypes are shared as descriptive text, workflow charts, and code books of medical concepts (all resources are available at https://www.phekb.org/). The Columbia eMERGE team converted PheKB code books by implementing the OMOP CDM[33], thus we obtained 20,640 condition concepts in 33 phenotypes. We excluded the concepts related to phenotype exclusion criteria. The number of concepts in each phenotype is provided in **Supplementary Material 1**.

Since the numbers of unique concepts trained in each of the four MCEs are different from each other, we constructed a standard evaluation set for each MCE to create a comparable evaluation. The standard evaluation set consists of the intersection between trainable concepts in the respective MCE and concepts from all phenotypes in PheKB (**Figure 2**). We excluded the out-of-bag concepts (i.e. concepts from PheKB phenotypes but not used to train the MCEs or concepts used to train MCEs but not in PheKB phenotypes) from the standard evaluation set. For example, if concept *C* appears in the knowledge graph but never appears in any of 33 phenotypes in PheKB, concept *C* was excluded from the standard evaluation set with regard to knowledge graph derived MCE since it is impossible to evaluate the embedding of the concept *C*.

In real-world applications, feature engineering in phenotyping is often started by generating seed concepts (i.e. seed features). Thus, we quantitatively assess different MCEs' performance in identifying phenotype-relevant concepts given varying number of seed concepts, from a single

seed concept to multiple seed concepts. We also assess interpretability of learned MCEs by plotting on 2-dimensional space.

Evaluation based on a single seed concept

Recall@k and *Precision@k* are commonly used metrics to assess how well the retrieved results satisfy a user's query intent in information retrieval. In our evaluation, the query was the embedding of the seed concept, and the retrieved results were candidate concepts. Given a single seed concept selected from a phenotype, each MCE retrieved the candidate concepts nearest to the seed concept. We used cosine similarity for the measure of distance between embeddings. Since the number of relevant concepts in a phenotype varies across the 33 phenotypes, we used a modified version of *Precision@k – Precision@k%* as our evaluation metric to provide a more consistent comparison across different phenotypes. Specifically, *Precision@k%* for phenotype *p* based on MCE e is defined as Eq (3):

$$Precision@k\%_{p,e} = \frac{1}{t}\sum_{C_i \in p} \frac{TP_{C_i}@k\%}{k\% * t} \qquad Eq(3)$$

where t is the number of unique concepts in phenotype p, TP_{C_i} is the number of relevant concepts retrieved for the embedding of the seed concept C_i, and $k\%$ is the percentage that controls the number of nearest concepts retrieved for the seed concept based on the cosine similarity. Similarly, *Recall@k%* is defined as Eq (4):

$$Recall@k\%_{p,e} = \frac{1}{t}\sum_{C_i \in p} \frac{TP_{C_i}@k\%}{t} \qquad Eq(4)$$

Note that for a fixed phenotype and MCE, $Recall@k\% = Precision@k\% * k\%$. We report the average *Precision@k%* and *Recall@k%* for each MCE by averaging the results from all phenotypes.

Evaluation based on multiple seed concepts

We again evaluated the *Precision@k%* and *Recall@k%*, except this time using multiple concepts selected from a phenotype for seeding the recommendations in each iteration. Given *n* seed concepts derived from a phenotype, the embedding for the seed was obtained by summing the embeddings of all *n* seed concepts, then each MCE retrieved the candidate concepts nearest to the seed embedding. Seed concepts for a phenotype were randomly chosen from the set of concepts in the phenotype. The number of seed embeddings was set to the number of unique concepts in the phenotype, yielding the same number of seed embeddings as in the single concept seed case. We performed this evaluation with *n* = 5 and *n* = 10. Similar to before, *Precision@k%* and *Recall@k%* for phenotype *p* based on MCE *e* are defined as Eq (5) and Eq (6) respectively:

$$Precision@k\% = \frac{1}{t}\sum_{i=1}^{t} \frac{TP_{seed_i}@k\%}{k\% * t} \qquad Eq(5)$$

$$Recall@k\% = \frac{1}{n}\sum_{i=1}^{t} \frac{TP_{seed_i}@k\%}{t} \qquad Eq(6)$$

where t is the number of unique concepts in phenotype p, TP_{seed_i} is the number of relevant concepts retrieved for the embedding of the $seed_i$, and $k\%$ is the percentage that controls the number of nearest concepts retrieved for the seed concept based on the cosine similarity. We report average *Precision@k%* and *Recall@k%* for each MCE by averaging the results from all phenotypes.

Visualization of learned embeddings

To assess interpretability of the learned embeddings, we plotted the embeddings of 1,000 randomly selected concepts from the intersection between the evaluation set of all MCEs on 2-dimensional space using t-SNE[34,35]. For clear visualization, we excluded concepts that were contained in multiple phenotypes in the process of selection of concepts.

RESULTS

Overall Performance

In total, there are 5204, 5204, 3636, and 3619 concepts in the evaluation sets for *graphEmb*, *graphEmb+*, *visitEmb* and *5yearEmb*, respectively. The average *Recall@k%* and *Precision@k%* of all MCEs based on a single seed concept are shown in **Figure 3**. **Figure 4** shows the average *Recall@k%* and *Precision@k%* of all MCEs based on 5 and 10 seed concepts, respectively. In both single seed concept and multiple seed concepts scenarios, graph-based MCEs outperformed co-occurrence-based MCEs. Specifically, *graphEmb+* outperformed other MCEs in average *Recall@k%* and *Precision@k%*.

Performance on Individual Phenotypes

Figure 5 shows *Recall@500%* of all 33 phenotypes for each MCE based on a single seed concept (**Fig. 5A**) and 5 seed concepts (**Fig. 5B**). *Recall@k%* and *Precision@k%* for all the other *k%* based on single and multiple seed concept(s) are provided in **Supplementary Material 2**. We excluded one phenotype (*Diverticulosis*), which contains less than 10 concepts, from the evaluation based on multiple seed concepts. In the evaluations based on a single concept seed and on five seed concepts, *graphEmb+* had the best performance among the MCEs in more than half of the evaluated phenotypes. Note that the results based on 10 seed concepts were similar to the results based on 5 seed concepts.

Visualization of the learned embeddings

t-SNE scatterplots of the 1,000 concepts for MCEs are shown in **Figure 6**. The color of each marker represents the phenotype that the concept belongs to, and the dashed circles are manual annotations indicating clusters of concepts belonging to the same phenotypes.

DISCUSSION

In this study, we assessed the potential of different MCEs for feature engineering in phenotyping. We evaluated four MCEs learned from two widely used resources, knowledge graphs and concept co-occurrences from EHR. The best *Precision@100%* is 0.35 from *graphEmb+*, indicating 1 out of 3 retrieved concepts is relevant when using MCEs for identifying phenotype-relevant concepts. Zhang et al. proposed a method for select phenotype relevant features in high-throughput phenotyping by mining the public literature[10]. Here we demonstrated MCE as an alternative approach for feature extraction in high-throughput phenotyping.

Figure 3 demonstrates that MCEs learned from knowledge graphs (*graphEmb* and *graphEmb+*) outperformed MCEs learned from co-occurrence statistics of EHR (*visitEmb* and *5yearEmb*), which may indicate that the current phenotyping efforts in the eMERGE network are more likely to focus on leveraging hierarchically structured clinical terminologies such as SNOMED-CT. Since most of the phenotypes developed in eMERGE network are rule-based, it is common for developers to navigate along the vocabularies to select the corresponding medical concepts.

Of the MCEs learned by using co-occurrence statistics of EHR, *5yearEmb* shows better performance than *visitEmb*. This is perhaps because phenotype-relevant condition concepts often appear across multiple visits and because there are irregular time intervals between visits. For example, concepts related to heart failure including initial presenting signs/symptoms and complications appear in multiple visits with the progression of heart failure. The 5-year window, which covers multiple visits, can capture sematic relationships between concepts with long-term relationships better than the visit window. *5yearEmb* also builds a less sparse co-occurrence matrix than *visitEmb*, which could help capture the underlying semantic relationships between medical concepts. *5yearEmb*, however, could introduce noise into the co-occurrence statistics for some acute diseases where intra-visit information between concepts are more important than inter-visit information. Therefore, if one aims to learn MCE by using co-occurrence statistics of EHR for identifying specific phenotype-relevant concepts, careful consideration of the characteristics of the phenotype might be required while selecting the window size. For example, visit-window can be used to learn MCE for phenotyping acute diseases such as clostridioides difficile and wider window (e.g., lifetime window or 5-year window in this study) can be used to learn MCE for phenotyping the diseases where symptoms appear in a long period of time such as heart failure and chronic kidney disease.

The window size of co-occurrence statistics also can be adjusted based on data quality and purpose of a study, although commonly chosen window sizes are visit and lifetime windows. We decided to use EHR data from 2013 to 2017 and to use visit window and 5-year window to ensure the quality of EHR[29]. In addition, EHR co-occurrence statistics are often specific to a local medical institution due to differences in local practices and the difficulties in patient data sharing. Co-occurrence statistics should be calculated across multiple institutions to learn more efficient and generalizable MCEs while minimizing bias of EHR from each institution[36-38]. The OMOP CDM can be leveraged to accomplish this since it provides a mechanism to convert local data models to its common data model.

The difference in results between *graphEmb* and *graphEmb+* show that enriching a knowledge graph by introducing additional relationships that have connection with the existing nodes in the knowledge graph is potentially beneficial for efficient learning of MCEs. This finding aligns well with the result from Shen et al.[39], where the authors obtained efficient embeddings for concepts in HPO using an enriched knowledge graph. It is not always true, however, that enriching a knowledge graph will lead to efficient learning of MCEs. For example, introducing a singular node that is connected to less than one existing node cannot lead to efficient learning since the singular node does not increase connectivity of the knowledge graph, which is critical for learning efficient embedding of the nodes. This limitation prevents MCE from learning based on a knowledge graph that is built upon the concepts from multiple distinct domains (e.g., condition and drug domains) that lack inter-domain relationships. The current version of OMOP CDM also has this limitation, since there are only 22,334 relationships between condition and drug concepts compared to 2,148,636 and 23,435,796 relationships between condition-condition and drug-drug concept pairs, respectively, in the *concept_relationship* table.

From **Figures 3** and **4**, we can see that all MCEs showed improved performance with the increasing number of seed concepts used to generate the seed embedding. In this study, the concepts in the multi-concept seeds were randomly selected; we hypothesize that performance can be improved with the use of fine-tuned and carefully selected seed concepts. However, as a trade-off for increased performance, providing a number of carefully selected seed concepts requires more effort from domain experts.

We can see from **Figure 6** that *graphEmb* and *graphEmb+* show better embeddings based on alignment with phenotypes than *visitEmb* and *5yearEmb*. From the figure, we observe a fewer number of clear clusters that align with phenotypes in *visitEmb* and *5yearEmb* than *graphEmb* and *graphEmb+*. This result suggests that co-occurrence information from EHR is not sufficient for learning interpretable embeddings that are consistent with phenotypes. It is interesting to see that other studies also found that simple co-occurrence information cannot learn interpretable representations that align with medical knowledge, although they did not use phenotyping knowledge to assess interpretability[21,40]. Interpretability of the embeddings, however, is not necessary for data-driven high-throughput phenotyping since the domain knowledge can be error-prone. Additionally, co-occurrence statistics derived from the EHR can reflect the daily clinical operation, providing complementary information to ontological knowledge during phenotyping development. In addition to the medical codes, many of the phenotypes developed in the eMERGE network required information from operational data elements (e.g., visit of a specific department)[3]. These data elements are not available in any medical vocabularies, but it can be achieved by using medical codes as proxies learned from the co-occurrence statistics derived from the EHR.

Representation learning including neural embedding is a rapidly evolving field. We acknowledge that besides the embedding methods investigated in this study, there are other more sophisticated methods. Although this study focused on evaluation of MCEs learned by using co-occurrence statistics of EHR and knowledge graphs, our evaluation framework can be generalized to wide range of MCEs learned by using diverse data sources and methods. Most recently, there have been several studies that combine diverse MCEs to learn representations containing richer information such as MMORE[41]. Future work will include learning enhanced MCE for phenotyping by combining multiple MCEs that show good performance in evaluation. Another limitation of this study is that the phenotypes used in the evaluation are mostly derived from rule-based approaches, making it difficult for us to evaluate the performance of different MCEs for data-driven phenotyping. Since learned MCEs, however, can naturally be served as inputs for machine learning based models, we expect that high-throughput phenotyping can leverage the evaluated MCEs in the future.

CONCLUSIONS

We assessed the potential of four different MCEs in feature engineering for phenotyping. MCEs learned by using knowledge graphs connected via hierarchical relationships between concepts outperformed MCEs learned by using co-occurrence statistics of EHR in identifying phenotype-relevant concepts. We also found that enriching a knowledge graph by adding relationships that increase connectivity of the knowledge graph improves MCE's performance in identifying phenotype-relevant concepts. Future works will include learning enhanced MCEs for phenotyping by combining multiple efficient MCEs and leveraging evaluated MCEs for data-driven high-throughput phenotyping.

CONTRIBUTORS

JL and CL implemented the methods and conducted all the experiments. JHK and AB contributed to conducting experiments and evaluation of the results. NS, CP, KN, and PR contributed to generating dataset and implementing OMOP CDM for PheKB phenotypes. CT and CW co-supervised the research and edited the manuscript. All authors were involved in developing the ideas, drafting and finalizing the paper.

FUNDING

This work was supported by National Library of Medicine grants R01LM009886 and 1R01LM012895-03, National Human Genome Research Institute grant U01HG008680, and National Center for Advancing Translational Science grant 1OT2TR003434-01.

COMPETING INTERESTS

The authors have no competing interests to declare.

SUPPLEMENTARY MATERIAL

Supplementary materials are available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGEMENT

We would like to thank eMERGE phenotyping workgroup who inspired this study.

REFERENCES

1. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association.* 2013;20(1):117-121.
2. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annual review of biomedical data science.* 2018;1:53-68.
3. Shang N, Liu C, Rasmussen LV, et al. Making work visible for electronic phenotype implementation: Lessons learned from the eMERGE network. *J Biomed Inform.* 2019;99:103293.
4. Wei W-Q, Leibson CL, Ransom JE, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *Journal of the American Medical Informatics Association.* 2012;19(2):219-224.
5. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *Journal of the American Medical Informatics Association.* 2012;19(2):212-218.
6. Yu S, Liao KP, Shaw SY, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc.* 2015;22(5):993-1000.
7. McCoy TH, Jr., Yu S, Hart KL, et al. High Throughput Phenotyping for Dimensional Psychopathology in Electronic Health Records. *Biol Psychiatry.* 2018;83(12):997-1004.
8. Gronsbell J, Minnier J, Yu S, Liao K, Cai T. Automated feature selection of predictors in electronic medical records data. *Biometrics.* 2019;75(1):268-277.
9. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat Protoc.* 2019;14(12):3426-3444.

10. Liao KP, Sun J, Cai TA, et al. High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. *J Am Med Inform Assoc.* 2019;26(11):1255-1262.
11. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence.* 2013;35(8):1798-1828.
12. Weng W-H, Szolovits P. Representation Learning for Electronic Health Records. *arXiv preprint arXiv:190909248.* 2019.
13. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association.* 2018;25(10):1419-1428.
14. Beam AL, Kompa B, Fried I, et al. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. April 2018. In:2019.
15. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep.* 2016;6:26094.
16. Bai T, Chanda AK, Egleston BL, Vucetic S. Joint Learning of Representations of Medical Concepts and Words from EHR Data. *Ieee Int C Bioinform.* 2017:764-769.
17. Camacho-Collados J, Pilehvar MT. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research.* 2018;63:743-788.
18. Duch W, Matykiewicz P, Pestian J. Neurolinguistic approach to vector representation of medical concepts. *Ieee Ijcnn.* 2007:3115-+.
19. Kwiatkowska M, Michalik K, Kielan K. Computational Representation of Medical Concepts: A Semiotic and Fuzzy Logic Approach. *Stud Fuzz Soft Comp.* 2012;273:401-420.
20. Lamy JB, Duclos C, Bar-Hen A, Ouvrard P, Venot A. An iconic language for the graphical representation of medical concepts. *Bmc Med Inform Decis.* 2008;8.
21. Choi E, Bahadori MT, Song L, Stewart WF, Sun J. GRAM: graph-based attention model for healthcare representation learning. Paper presented at: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining2017.
22. Shen F, Peng S, Fan Y, et al. HPO2Vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the Human Phenotype Ontology. *J Biomed Inform.* 2019;96:103246.
23. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. Paper presented at: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)2014.
24. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Paper presented at: Advances in neural information processing systems2013.
25. Choi E, Bahadori MT, Searles E, et al. Multi-layer representation learning for medical concepts. Paper presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining2016.
26. Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc.* 2016;23(6):1046-1052.
27. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Studies in health technology and informatics.* 2015;216:574.

28. *The Book of OHDSI.* Observational Health Data Sciences and Informatics; 2019.
29. Ta CN, Dumontier M, Hripcsak G, Tatonetti NP, Weng C. Columbia Open Health Data, clinical concept prevalence and co-occurrence from electronic health records. *Scientific data.* 2018;5:180273.
30. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. Paper presented at: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining2016.
31. Abadi M, Barham P, Chen J, et al. Tensorflow: A system for large-scale machine learning. Paper presented at: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)2016.
32. Zeiler MD. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:12125701*. 2012.
33. Hripcsak G, Shang N, Peissig PL, et al. Facilitating phenotype transfer using a common data model. *Journal of biomedical informatics.* 2019;96:103253.
34. Maaten Lvd, Hinton G. Visualizing data using t-SNE. *Journal of machine learning research.* 2008;9(Nov):2579-2605.
35. Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively. *Distill.* 2016;1(10):e2.
36. Lu C-L, Wang S, Ji Z, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *J Am Med Inform Assn.* 2015;22(6):1212-1219.
37. Tian Y, Shang Y, Tong D-Y, et al. POPCORN: A web service for individual PrognOsis prediction based on multi-center clinical data CollabORatioN without patient-level data sharing. *Journal of biomedical informatics.* 2018;86:1-14.
38. Tong J, Duan R, Li R, Scheuemie MJ, Moore JH, Chen Y. Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data. Paper presented at: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing2020.
39. Shen F, Peng S, Fan Y, et al. HPO2Vec+: leveraging heterogeneous knowledge resources to enrich node embeddings for the human phenotype ontology. *Journal of biomedical informatics.* 2019;96:103246.
40. Ma F, You Q, Xiao H, Chitta R, Zhou J, Gao J. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. Paper presented at: Proceedings of the 27th ACM International Conference on Information and Knowledge Management2018.
41. Song L, Cheong CW, Yin K, Cheung WK, CM B. Medical concept embedding with multiple ontological representations. Paper presented at: Proceedings of the 28th International Joint Conference on Artificial Intelligence2019.

Figure List

Figure 1. Application of visit window and 5-year window to patient records. Each mark along a patient's timeline (A) indicates a visit containing multiple concepts, C_x. Raw patient records (A) are transformed to the formats in (B) by applying a 5-year window and in (C) by applying visit windows.

Figure 2. Set diagrams between the unique concepts in PheKB and in each medical concept embedding (MCE). The intersection of each set diagram is the standard evaluation set for that MCE. Since we excluded EHRs that had less than two concepts in each window, there are slight difference in the total number of unique concepts between visit windowed and 5-year windowed EHR.

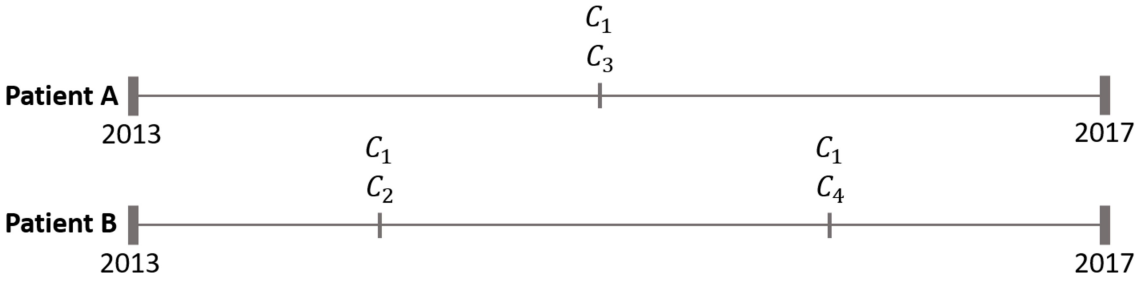
Figure 3. (A) *Recall@k%* and (B) *Precision@k%* of all medical concept embeddings (MCEs) based on a single seed concept.

Figure 4. Average *Recall@k%* and *Precision@k%* of all MCEs based on 5 and 10 seed concepts. (A) *Recall@k%* with 5 seed concepts. (B) *Precision@k%* with 5 seed concepts. (C) *Recall@k%* with 10 seed concepts. (D) *Precision@k%* with 10 seed concepts.

Figure 5. Average *Recall@500%* of all individual phenotypes based on (A) a single seed concept and (B) five seed concepts for MCEs. Full names for abbreviated phenotypes are provided in **Supplementary Material 1**.

Figure 6. t-SNE scatterplots of 1,000 randomly selected condition concepts from (A) *visitEmb*, (B) *5yearEmb*, (C) *graphEmb*, and (D) *graphEmb+*. Full names for abbreviated phenotypes are provided in **Supplementary Material 1**.

A

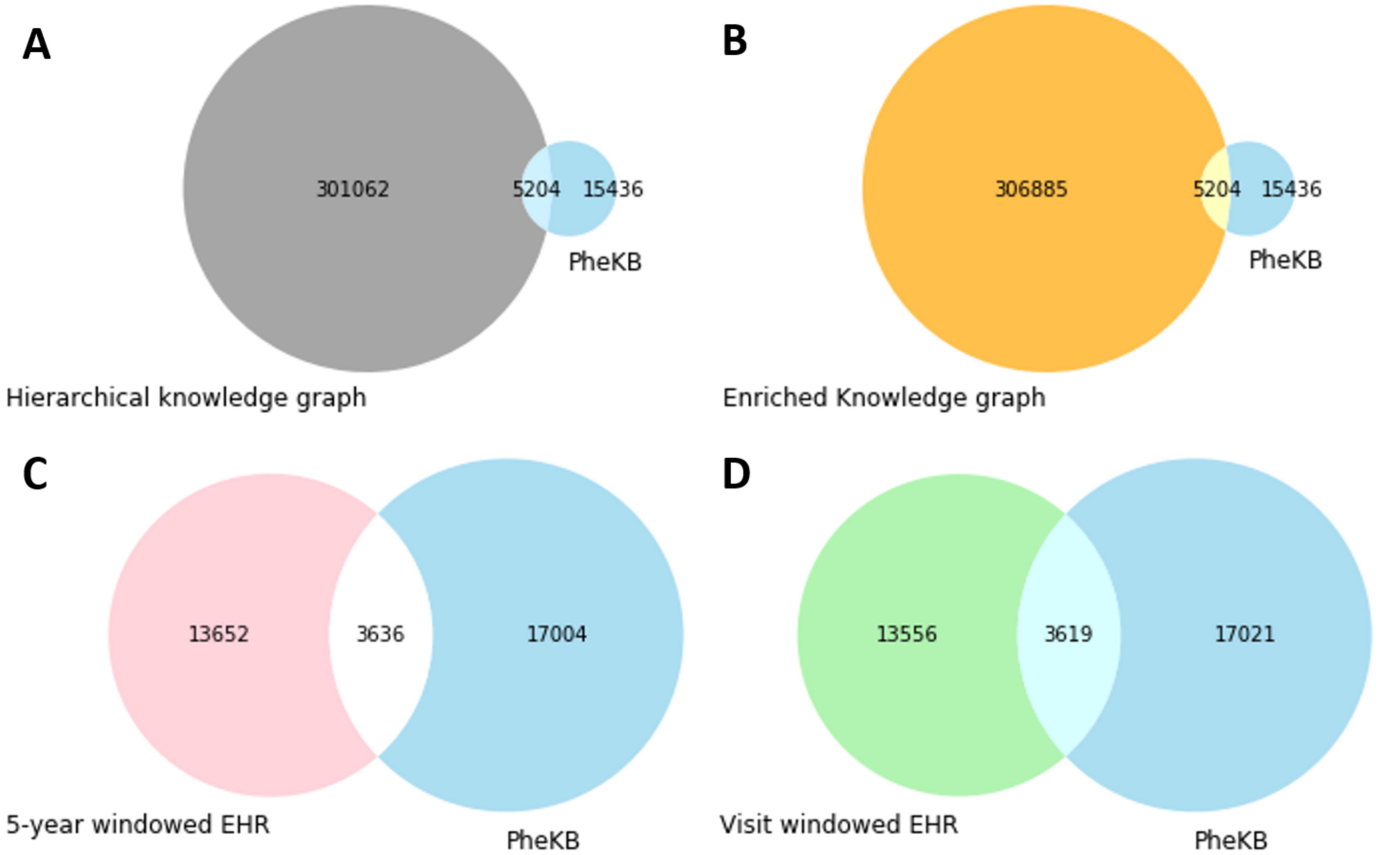


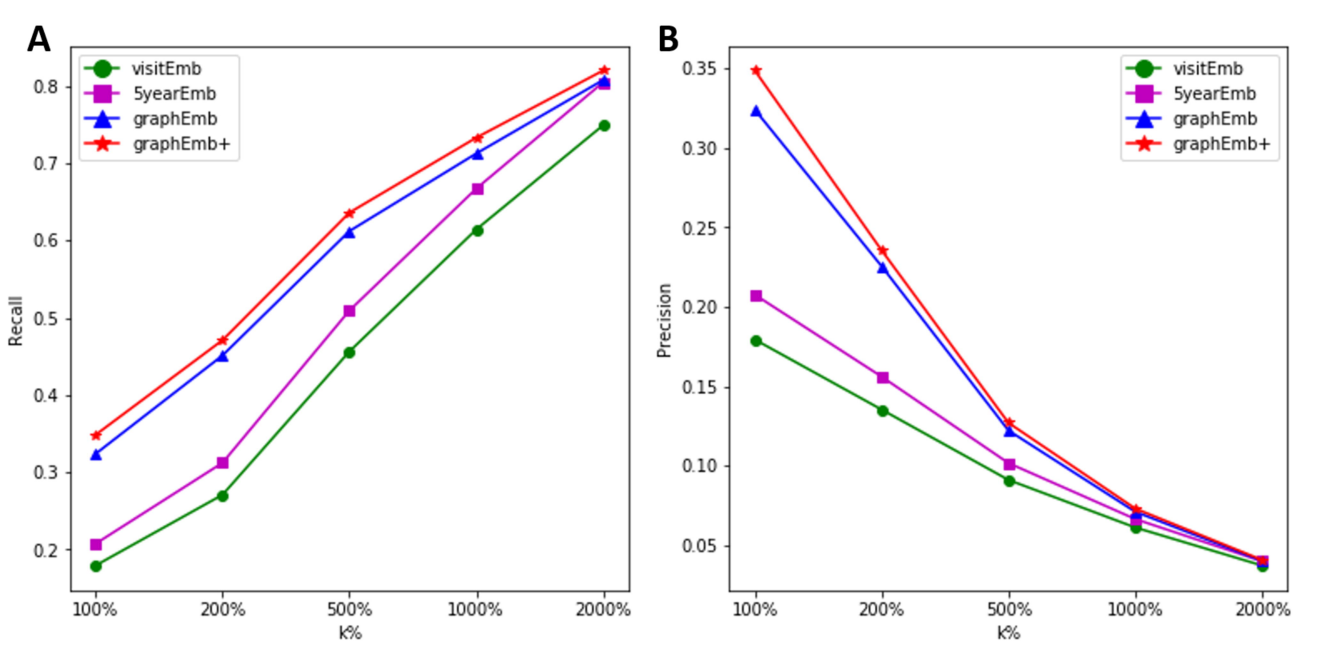
B

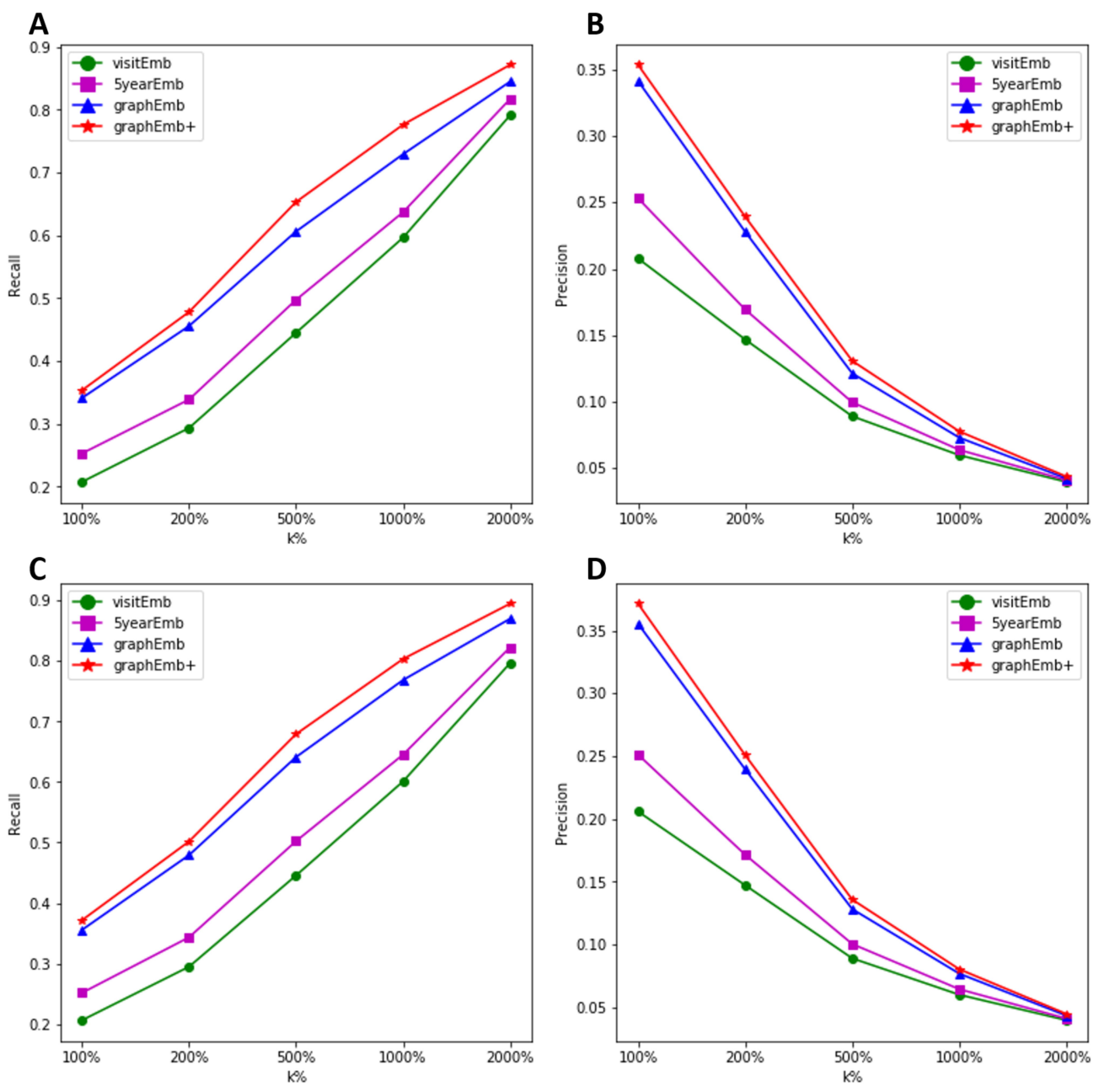
	Concepts
Record 1	C_1, C_3
Record 2	C_1, C_2, C_4

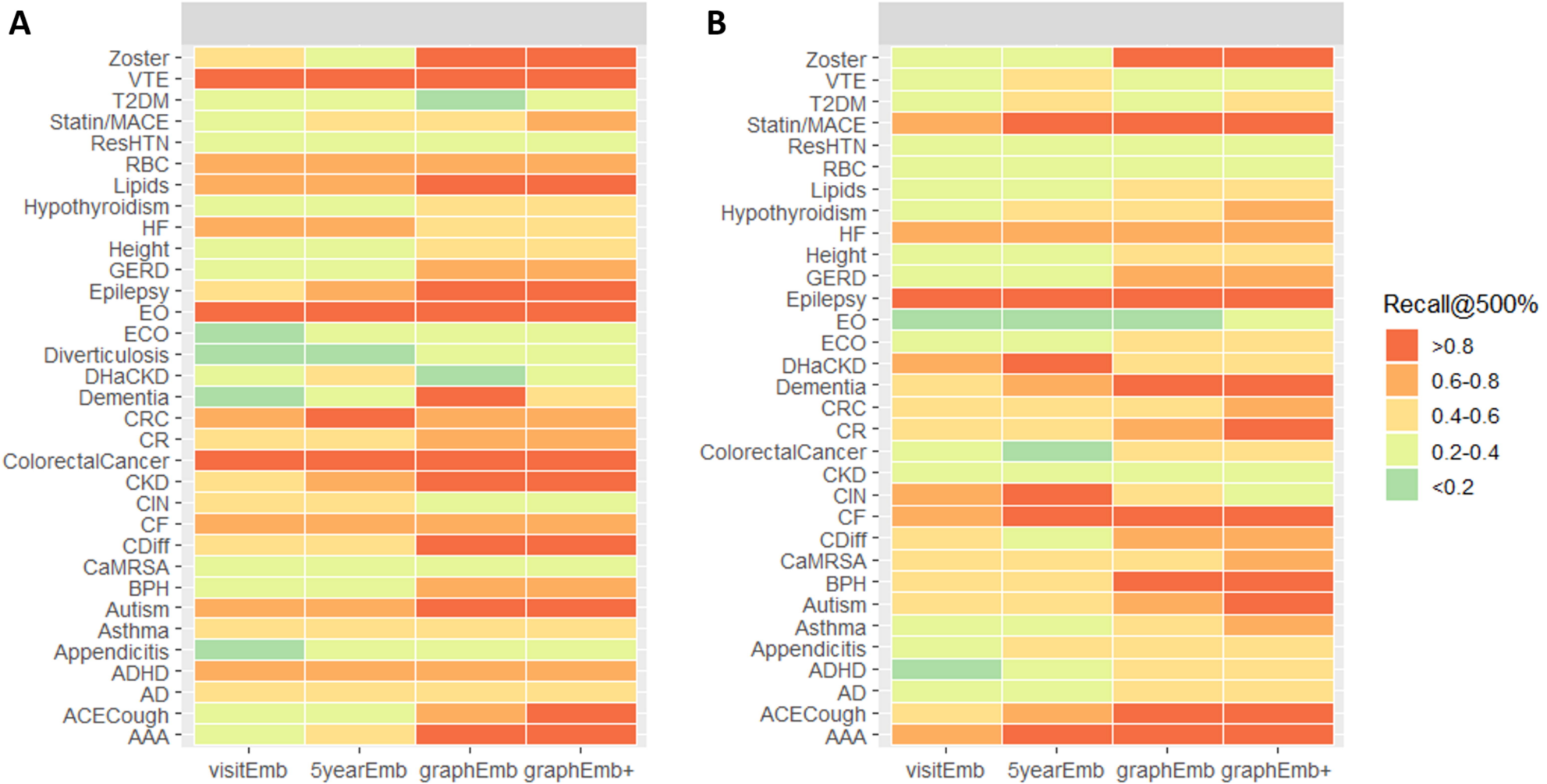
C

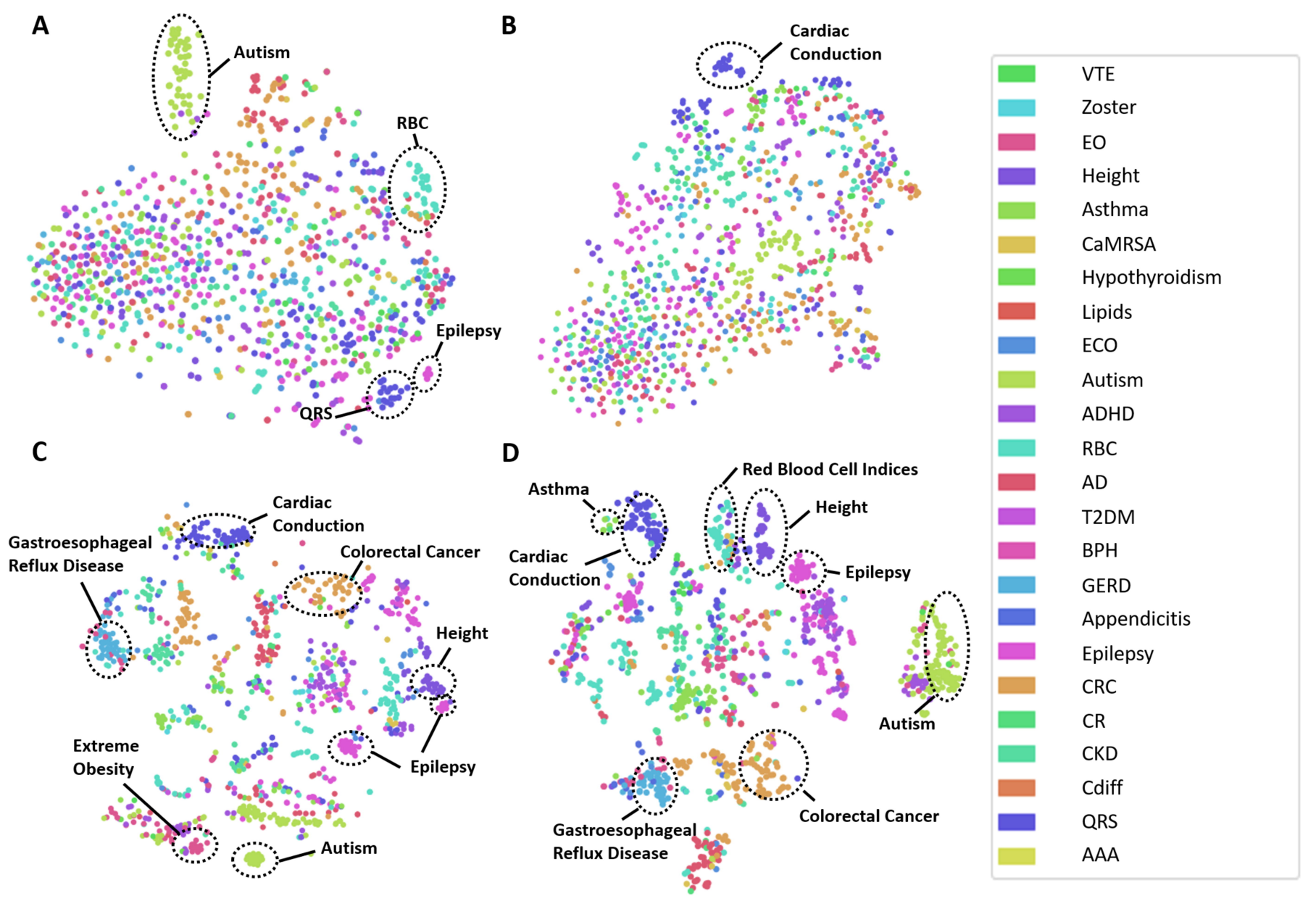
	Concepts
Record 1	C_1, C_3
Record 2	C_1, C_2
Record 3	C_1, C_4


A
301062
5204
15436
PheKB
Hierarchical knowledge graph
B
306885
5204
15436
PheKB
Enriched Knowledge graph
C
13652
3636
17004
5-year windowed EHR
PheKB
D
13556
3619
17021
Visit windowed EHR
PheKB


A
visitEmb
5yearEmb
graphEmb
graphEmb+
Recall
0.8
0.7
0.6
0.5
0.4
0.3
0.2
100%
200%
500%
1000%
2000%
k%
B
visitEmb
5yearEmb
graphEmb
graphEmb+
Precision
0.35
0.30
0.25
0.20
0.15
0.10
0.05
100%
200%
500%
1000%
2000%
k%


A
B
C
D
visitEmb
5yearEmb
graphEmb
graphEmb+
Recall
Precision
k%
100%
200%
500%
1000%
2000%
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.05
0.10
0.15
0.20
0.25
0.30
0.35


A
B
Zoster
VTE
T2DM
Statin/MACE
ResHTN
RBC
Lipids
Hypothyroidism
HF
Height
GERD
Epilepsy
EO
ECO
Diverticulosis
DHaCKD
Dementia
CRC
CR
ColorectalCancer
CKD
CIN
CF
CDiff
CaMRSA
BPH
Autism
Asthma
Appendicitis
ADHD
AD
ACECough
AAA
visitEmb
5yearEmb
graphEmb
graphEmb+
Recall@500%
>0.8
0.6-0.8
0.4-0.6
0.2-0.4
<0.2


A
Autism
RBC
Epilepsy
QRS
B
Cardiac
Conduction
C
Cardiac
Conduction
Gastroesophageal
Reflux Disease
Colorectal Cancer
Height
Extreme
Obesity
Epilepsy
Autism
D
Red Blood Cell Indices
Asthma
Height
Cardiac
Conduction
Epilepsy
Autism
Colorectal Cancer
Gastroesophageal
Reflux Disease
VTE
Zoster
EO
Height
Asthma
CaMRSA
Hypothyroidism
Lipids
ECO
Autism
ADHD
RBC
AD
T2DM
BPH
GERD
Appendicitis
Epilepsy
CRC
CR
CKD
Cdiff
QRS
AAA