# Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data

Link: <https://www.nature.com/articles/s41746-021-00519-z>

**Motivation**

What is the problem being solved?

* Develop multi-center large-scale code embeddings efficiently identify relevant features related to a disease of interest
* Integrate the different types of structured data efficiently (codified EHR data)
* Establish a highly cooperative and shareable clinical knowledge network across institutions

Why is it important?

* comprehensive longitudinal data allow for studies to examine a broad range of hypotheses
* Provide information about the dependency structure across different EHR elements -> provide information about the relationship of conditions and codes at a particular institution and equivalent codes across institutions.
* No longer be associated with individual patient data -> be readily shared -> facilitate multi-center collaborations

What previous work exists?

1. Manually select individual EHR features and map specific EHR codes to represent each feature, requiring input from clinical and informatics experts

* ICD (International Classification of Disease): disease conditions
* LOINC (Logical Observation Identifiers Names and Codes): laboratory tests
* CPT (Current Procedural Terminology) and CCS (Clinical Classifications Software): procedures
* RxNorm and NDC (National Drug Code): medications

1. Previous embeddings

* Code2Vec19
* Med2Vec21

Why is the previous work insufficient to solve the problem?

* Susceptible to subjective bias and time-consuming
* Sharing algorithms across institutions -> need to identify institution-specific codes and coding patterns in collaborative or replication studies.
* Require the use of patient-level data -> hinders the ability to share data across institutions.
* Previous embeddings were primarily derived for specific applications (not for the creation of knowledge networks) -> tuned the key hyper-parameters -> limit the applicability of the learned embedding vectors to other downstream tasks

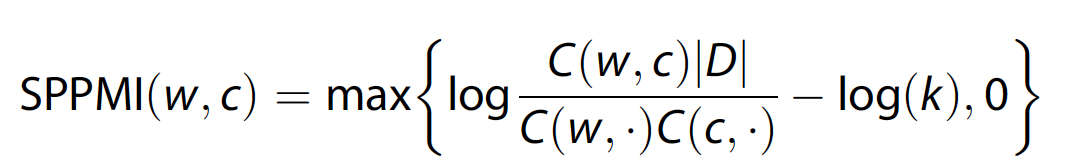
**Approach**

1. Data pre-processing

* EHR data from two large hospital systems
  + VA Corporate Data Warehouse (CDW)
  + Mass General Brigham (MGB).
* Aggregate four codified data domains (i.e., diagnosis, procedures, lab measurements, and medications) into PheCode, CCS, RxNorm, LOINC codes, and manual lab concepts
* Apply frequency control

1. Create embedding vectors (SVD-SPPMI algorithm)

* Co-occurrence matrix (C(w,c)):
  + For any given patient, scan through each of their codes as a target code. For any given target code occurring at time t (w\_t)
  + Count all codes occurring within 30 days of t as co-occurrences with w\_t
  + Total numbers of co-occurrences for all possible pairs of codes are aggregated over all target codes (i) within each patient and (ii) across all patients
* Constructed two separate co occurrence matrices at each site -> to select tuning parameters

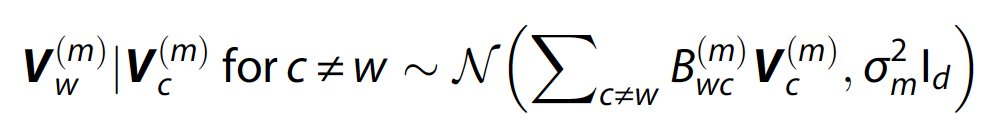


Where k=1; C(w, .)= row of sum of C(w,c); |D| = total sum of co-occurrence

* SPPMI matrix first d-dimensional SVD = U\_d diag(lamda\_1, lamda\_2,...)U\_d^T
* d-dimensional embedding vectors Vd = U\_d diag(sqrt(lamda\_1), sqrt(lamda)2,...)

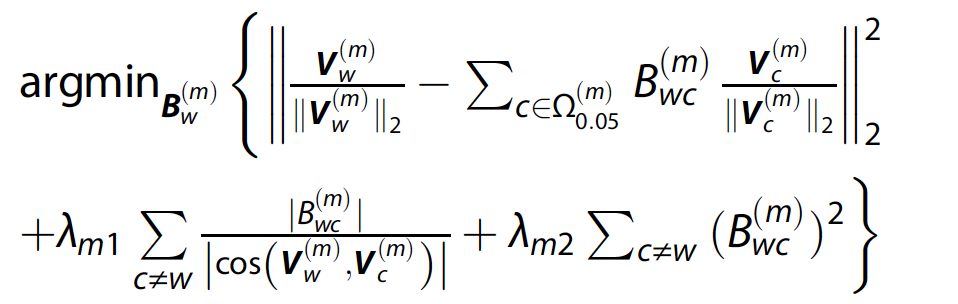
1. Feature selection

* Model conditional distribution of the target code w and other codes by imposing a Gaussian distribution on the embedding vectors and inferring the dependency via a Gaussian graphical model

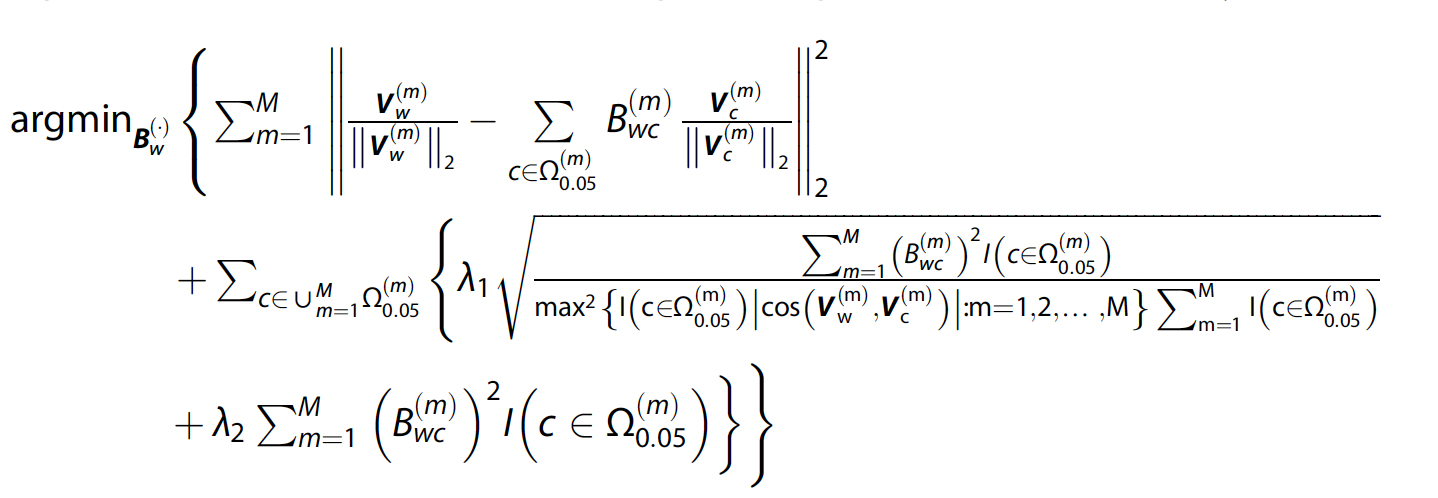


* B^(m)\_wc encodes the conditional dependency between codes w and c
  + B^(m)\_wc = 0 if V^(m)\_w is independent of V^(m)\_c given all other code vectors
* KESER (<https://celehs.connect.hms.harvard.edu/kesernetwork/>)
  + **Initial screening:** based on marginal cosine similarity and consider codes in Ω^(m)\_0:05 = {c : (cos(V^(m)\_w, V^(m)\_c) > = ρ0:05; c≠w}
  + **Single site: Elastic net penalized regression**

Estimate B^(m) = {B^(m)\_wc, c≠ w} as



* + **Multiple sites: Integrative least squared regression with a mixture of ridge and group sparse penalty**



* + B^(m)\_wc ≠ 0 -> important features
* Tuning parameters (lamda) selected by CV on SPPMI matrix

1. Evaluation of embeddings performance

* Accuracy in detecting known similar (if they shared the same integer to calculate the SNR and AUC) pairs and related pairs
  + AUC
  + sensitivity of detecting pairs by thresholding cosine similarities to achieve a FPR (Type I error) of 0.01, 0.05 or 0.10

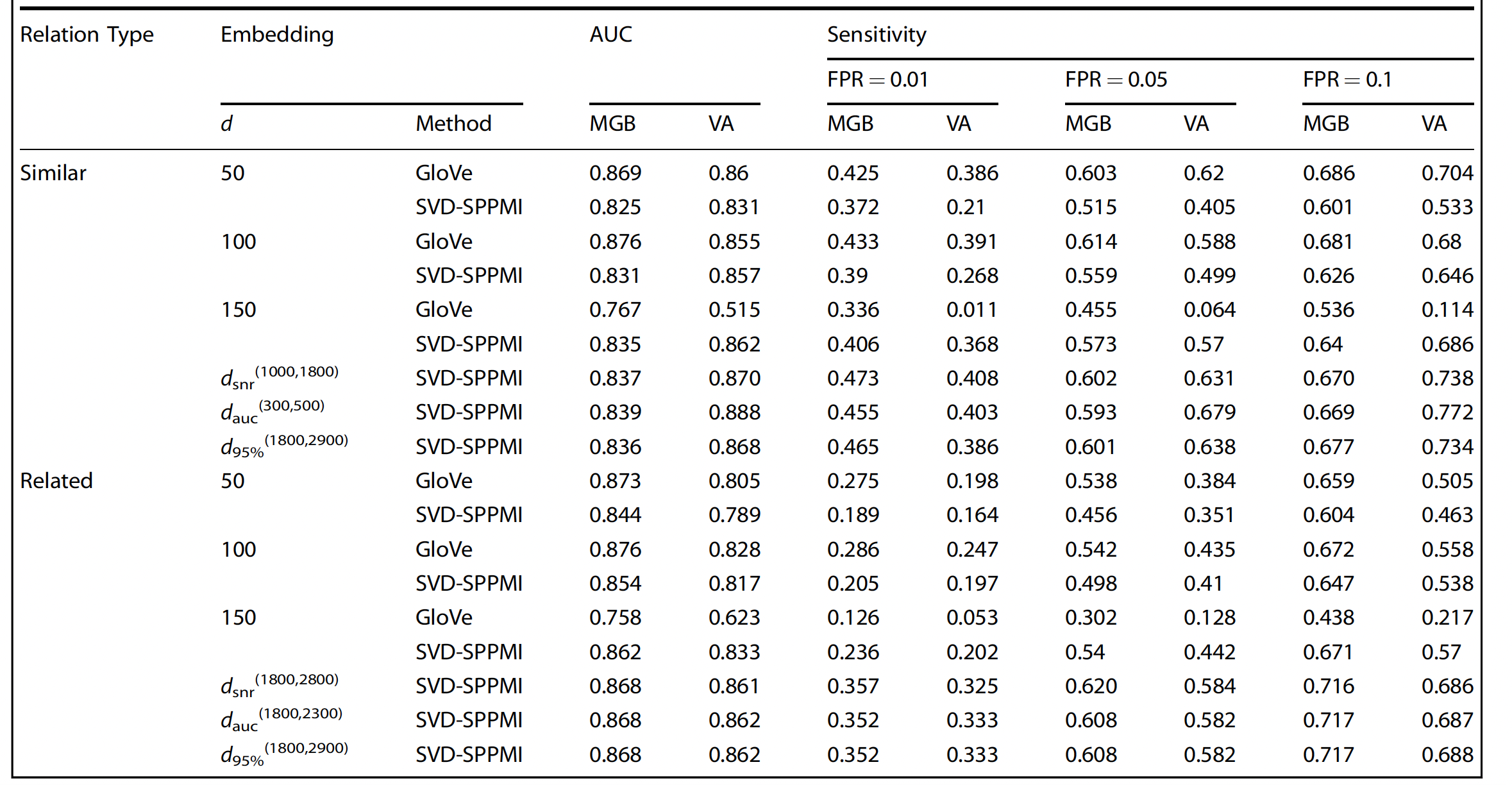
1. Knowledge mapping

* Trained embeddings at MGB and VA can be used to map codes across the two institutions via orthogonal transformation
* Find an orthogonal matrix Q to minimize the distance between V\_MGB and V\_VA\*Q

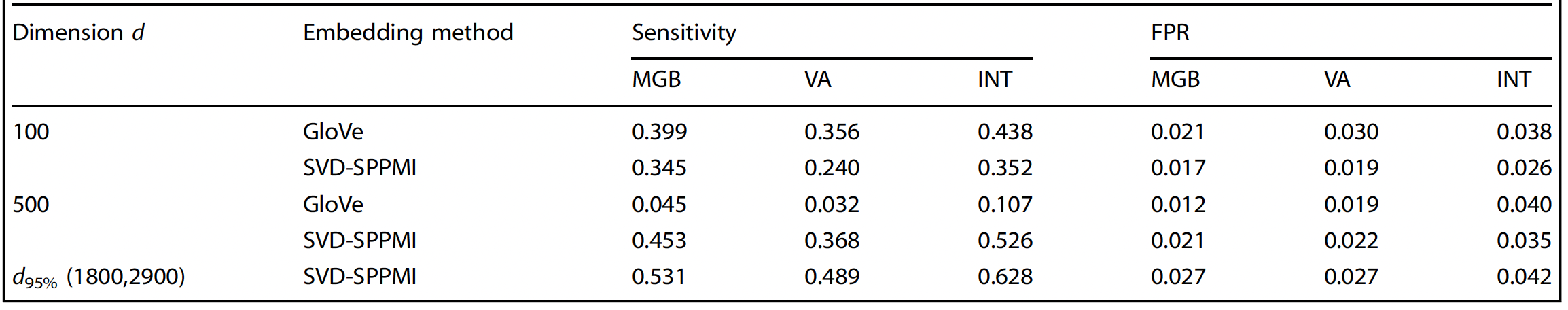
**Results**

1. Performance of embeddings pre-KESER

* Initial embedding dimensions: retaining 95% of the variation in the SVD (d\_95%) -> MGB: 1800; VA: 2900
* Choose the dimensions by maximizing either
  + the signal-to-noise ratio (SNR)
  + AUC associated with pairs with known relations against random pairs



1. Performance of embeddings post-KESER



1. Knowledge mapping

* 4 and 9 lab codes uniquely identified by VA and MGB respectively-> majority of the lab codes are unique to the site -> high cross-site heterogeneity in lab coding.
* KESERINT integrates data from both sites -> achieve higher accuracy in reflecting clinical knowledge
* Can successfully select informative and clinically meaningful features that can be used effectively for phenotyping and other downstream analyses

1. Evaluations

* Embedding algorithms: SVD-SPPMI > GloVe
* KESER: INT ~ MGB -> maybe MGB is sufficient
* Feature selection:
  + Averaged AUC: KESER\_MGB > PheCode, FULL, and SAFE features.
  + Averaged F-score: KESER\_MGB > PheCode, FULL, and SAFE features.
  + 95% CI of accuracies:
    - KESER-selected features ~ SAFE features
    - FULL features and main PheCode alone are substantially wider

**Contributions**

1. Facilitate multi-center collaborative studies using EHR data

* Past: require that each institution perform analyses individually with results compared across institutions
  + Coding behaviors, disease management and strategies, and healthcare delivery patterns can vary
* Opportunity to develop an integrated clinical knowledge network with input from many institutions
* Knowledge network can be updated over time across multiple healthcare systems

1. Allow the use of grouped lab codes in research studies

* Lab codes are much less standardized
* Past: no established grouping structure can be used at scale for research studies
* Grouped lab codes based on the similarity between the vectors

1. Tailor the dimension of embedding vectors to the goals of building knowledge network

* Past: GloVe 300; CUI 500
* Lower dim: identifying near synonymous concepts or translations
* Higher dim: assessing relatedness and embedding regression aiming to optimize feature selection and building knowledge networks
  + KESER regression performs better with higher dimensional embedding vectors

1. Computational efficient

* Past: SGNS
* SVD-SPPMI approach does not need to conduct the negative sampling

1. Robustness

* Past: GloVe has convergence issues when fitting for higher dimensions
* Sparsity of the SPPMI matrices

**Limitations**

* KESER high dimension embedding vectors are not suitable for all tasks

**Other**

1. Why does the method section come after the results and discussion section?
2. *“d\_95% is obtained by retaining 95% of the variation in the SVD.”* What does this mean?

Similar to variation in PCA

1. Same conditional distribution question as last week’s paper.
2. *“We evaluate the quality of the cross-institution mapping based on the top-1, top-5, and top-10 accuracy calculated based on the test set.”*

What are “top-1, top-5, and top-10 accuracy” referring to? I thought there was only 1 cross-institution map and 1 set of testing data. Then, how did they get multiple accuracy measures? Did they average out the accuracies of top-5/10?

machine learning top k: calculate accuracy based on top k observations that has highest prediction probability

<https://pub.aimind.so/learn-about-the-top-k-accuracy-score-using-the-glass-dataset-f2530c9c90ac>