Automated disease cohort selection using word embeddings improves upon electronic phenotyping algorithms

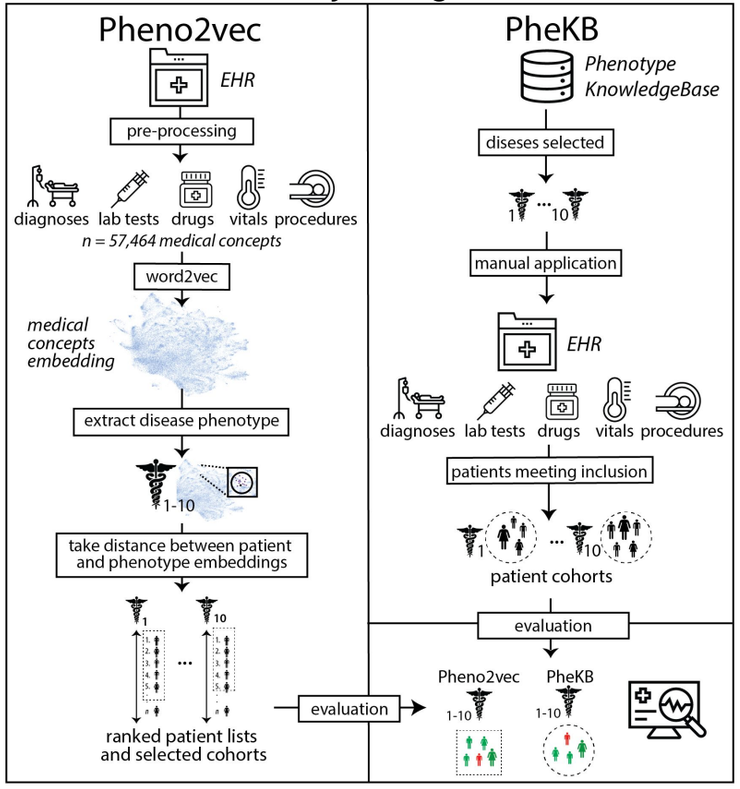
Jessica K. De Freitas, Benjamin S. Glicksberg, Kipp W. Johnson, Joel T. Dudley, Riccardo Miotto

[1]Hasso Plattner Institute for Digital Health at Mount Sinai; [2]Department of Genetics and Genomic Sciences Mount Sinai

Abstract

Here we present, Phe2vec, an unsupervised learning framework based on neural networks for EHR-based phenotyping, which aims to be scalable, robust and interpretable. Phe2vec derives vector-based representations, i.e., embeddings, of medical concepts to define the disease phenotypes using the semantic closeness in the embedding space to a seed concept (e.g., and ICD code) Embeddings are then aggregated at the patient-level to identify populations related to a specific disease based on distance with the phenotype in the embedding space. Experiments based on manual chart review show that Phe2vec performs, at least, as good as rule-based PheKB algorithms for different and diverse diseases and is capable of seamlessly creating reliable phenotypes also for diseases not covered by PheKB. To the best of our knowledge, this is the first time a head-to-head comparison between an automated phenotyping method and current gold-standard rule-based algorithms has been assessed. Being based on unsupervised embeddings of medical concepts, Phe2vec can also potentially be leveraged as the first layer of clinical predictive learning systems and can be elegantly extended to include other modalities of data, leading to phenotypes related to an holistic view of the diseases.

Study Design



Methods

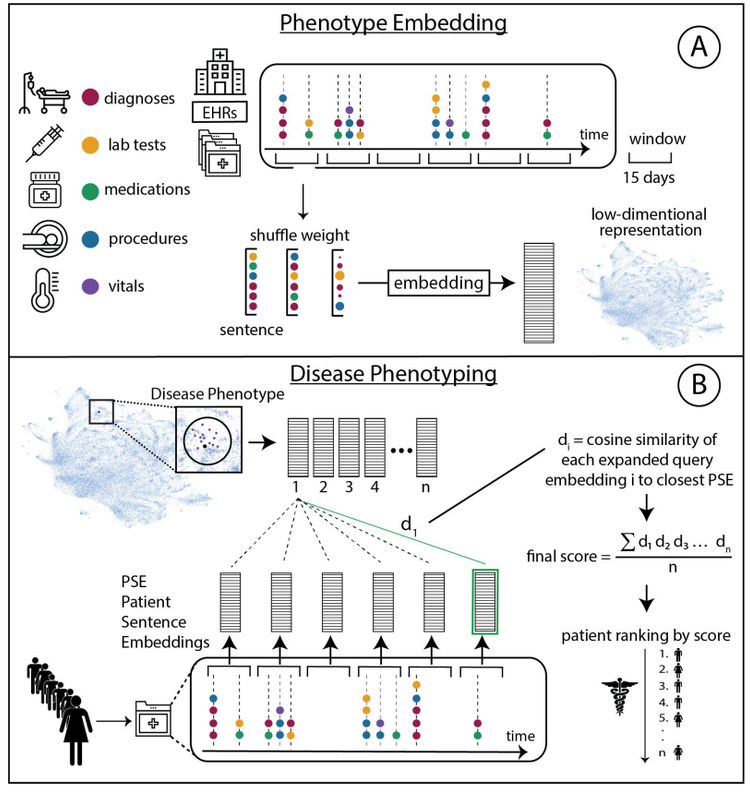


Figure 2: Phe2vec framework. (A) Embedding algorithm to create low-dimensional vector-based representations of medical concepts from longitudinal EHRs. (B) Disease phenotypes are defined considering a "seed" concept (e.g., an ICD code) and its neighbors in the embedding space. A patient clinical history is summarized aggregating all the medical concept embeddings. This representation is then used to measure the closeness with the phenotype in the metric space to determine his/her association to a disease.

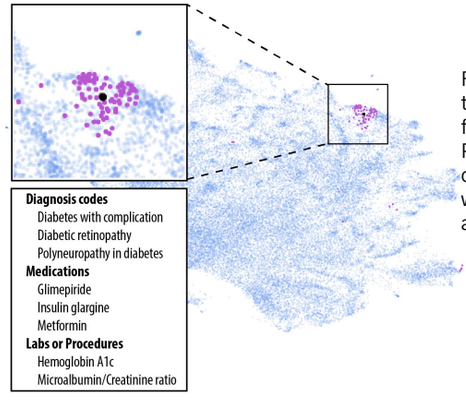


Figure 3: UMAP visualization of the EHR-based phenotype space for Type 2 Diabetes generated by Phe2vec with word2vec. Seed concepts are colored in black, while concepts in the phenotypes are colored in purple.

Results

		F-score	R-precision	AUC-PR
Seed Code	BoCon	0.42	0.39	0.53
	Word2vec	0.50	0.52	0.59
	GloVe	0.43	0.51	0.54
	Fasttext	0.47	0.50	0.58
Disease Phenotype	BoCon	0.53	0.44	0.56
	Word2vec	**0.64**	**0.62**	**0.69**
	GloVe	0.57	0.53	0.62
	Fasttext	0.59	0.54	0.64

Table 1: Disease cohort selection results obtained with the automated evaluation, where PheKB cohorts are considered as gold standard. We compare embedding-based methods (Word2vec, GloVE, Fasttext), which rely on similarity between patients and phenotypes, and bag of codes (BoC), which just count the frequency of the phenotype concepts in the patient history. All results are average across ten diseases.

Disease	Phe2vec	PheKB
Abdominal aortic aneurysm	1.00	0.95
Attention deficit hyperactivity disorder	0.97	0.85
Atrial Fibrillation	0.85	0.85
Autism	0.81	0.95
Crohn's disease	0.98	0.81
Dementia	0.98	0.82
Herpes zoster	0.92	0.45
Multiple sclerosis	0.97	0.97
Sickle cell disease	0.96	0.83
Type 2 diabetes mellitus	0.98	0.74

Table 2: Positive predictive value obtained by Phe2vec and PheKB against gold standard via manual chart review of progress notes

Conclusions

We developed and validated an architecture named Phe2vec that infers informative vector-based representations of medical concepts and uses distance analysis from a seed concept to define phenotypes to retrieve cohorts of patients associated with diseases. Experiments highlight the slight prevalence of word2vec as a model to learn medical concepts embeddings form EHRs over GloVe and FastText. Phe2vec aims to be domain-free, robust and scalable to all diseases. Experimental results on large-scale EHRs show that Phe2vec identifies similar phenotypes to PheKB and can be used to accurately identify disease cohorts. In particular, Phe2vec performs on par or outperforms PheKB algorithms in nine out of ten selected diseases.

HPI·MS