New results for "Unsupervised Patient Phenotyping for Large-Scale EHR via Distributed Bi-Factored Inference"

1 Detecting Known-relationship Pairs

type	subtype	num	DIANE	Bert	BioBert	PubmedBert	SAPBert
similar pairs	RxNorm Hierachy	650	**0.799**	0.498	0.569	0.601	0.656
	PheCode Hierachy	510	**0.937**	0.638	0.584	0.605	0.830
	LOINC Hierachy	477	0.610	0.910	0.856	0.857	**0.984**
related pairs	Differential Diagnosis	1226	**0.768**	0.590	0.562	0.565	0.725
	May Treat	1133	0.747	0.653	0.616	0.656	**0.761**
	Classifies	751	**0.850**	0.545	0.509	0.545	0.730
	May Prevent	171	**0.802**	0.602	0.578	0.651	0.785

Table 1: AUC of detecting different types of know-relationship pairs with estimated model parameter $\widehat{\Theta}$ and the number of pairs within each subtype.

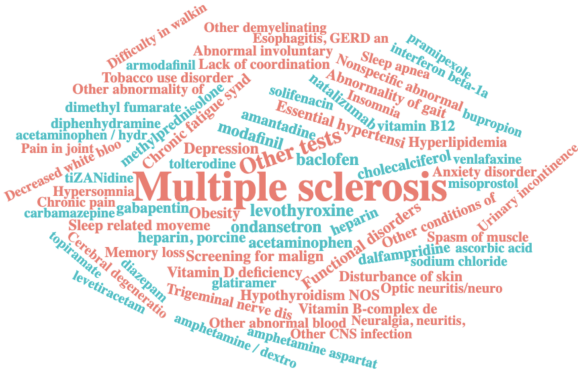
2 Risk Prediction

	DIANE	Bert	BioBert	PubmedBert	SAPBert
AUC	**0.854**	0.770	0.691	0.740	0.770
Brier score	**0.049**	0.069	0.073	0.072	0.067

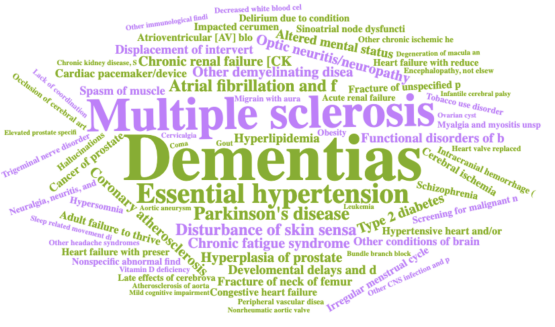
Table 2: The risk prediction performance of decision trees with patient embeddings obtained from different methods, which is evaluated by the AUC and Brier score. The model is trained with a half patients and evaluated on the rest of the patients.

3 Patient Phenotyping

(a) top features of patients with AD



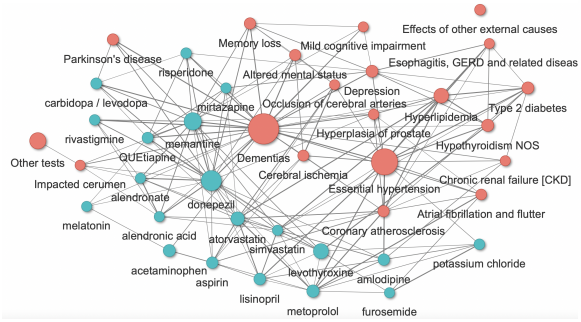
(b) top feaetures of patients with MS



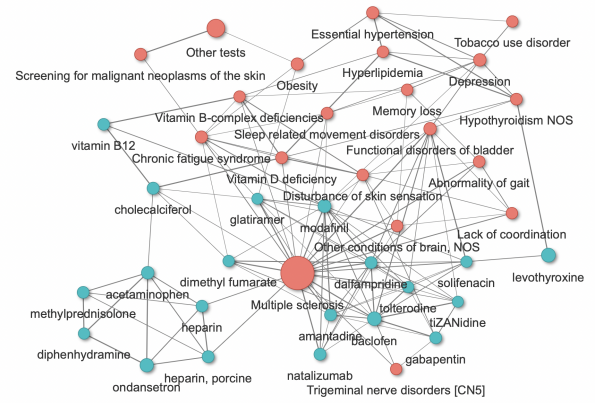
(c) phenotype features

(d) drug features

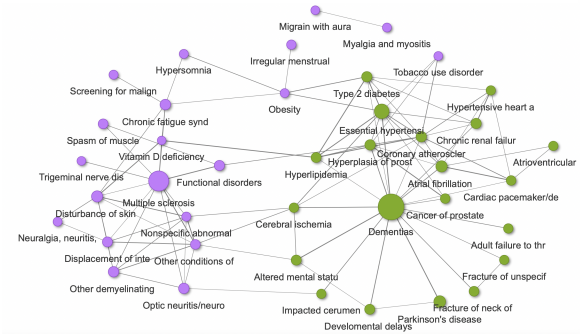
Figure 1: The word clouds of (a) top features of AD patient; (b) top features of MS patient; (c) phenotype features that drive the differences between AD patient and MS patient and (d) drug features that drive the differences between AD patient and MS patient. The size of the feature is determined by the occurrence probability in figure (a)(b) and between-group difference in figure (c)(d). In figure (a)(b), red colored features are disease code (PheCode) while blue colored features are drug code (RxNorm). In figure (c)(d), green colored features represent higher average intensity in the AD patient and purple colored features represent higher intensity in the MS patient.



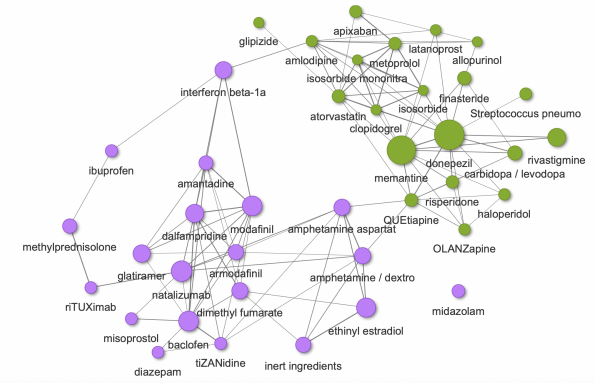
(a) top features of patients with AD



(b) top feaetures of patients with MS



(c) phenotype features



(d) drug features

Figure 2: The Networks of (a) top features of AD patient; (b) top features of MS patient; (c) phenotype features that drive the differences between AD patient and MS patient and (d) drug features that drive the differences between AD patient and MS patient. The size of the node is determined by the occurrence probability in figure (a)(b) and between-group difference in figure (c)(d). In figure (a)(b), red colored features are disease code (PheCode) while blue colored features are drug code (RxNorm). In figure (c)(d), green colored features represent higher average intensity in the AD patient and purple colored features represent higher intensity in the MS patient. The width of the edge is determined by the estimated parameter $\widehat{\mathbf{\Theta}}$.