Word embeddings trained on published case reports are lightweight, effective for clinical tasks, and free of protected health information

Zachary N. Flamholz[1], Andrew Crane-Droesch, PhD[2,3], Lyle H. Ungar, PhD[4,5], Gary E. Weissman, MD, MSHP[3,5,6,7]

[1]Medical Science Training Program, Albert Einstein College of Medicine, [2]Penn Medicine Predictive Healthcare. University of Pennsylvania Health System, [3]Palliative and Advanced Illness Research (PAIR) Center, University of Pennsylvania Perelman School of Medicine , [4]Department of Computer and Information Science, University of Pennsylvania, [5]Institute for Biomedical Informatics, University of Pennsylvania , [6]Leonard Davis Institute of Health Economics, University of Pennsylvania, [7]Pulmonary, Allergy, and Critical Care Division, University of Pennsylvania Perelman School of Medicine

EINSTEIN
Albert Einstein College of Medicine

Background

- Clinical encounter notes represent a potentially rich source of patient information for use in clinical predictive algorithms
- Word embeddings are low-dimensional, vector representations of language learned from a text corpus that can facilitate downstream prognostic and classification tasks in the clinical domain
- Previous studies utilize embeddings that are trained on either general language corpora that may not adequately represent clinical language or text containing patient protected health information that cannot be made public

Objective

- Develop and make publicly available a fully de-identified, clinical language embedding model from published case reports
- Evaluate word embeddings trained on a variety of text corpora for the ability to predict in-patient mortality from the text of physician encounter notes

Methods

- We trained 60 word embeddings using a combination of 3 training algorithms, 5 text corpora, and 4 vector dimensions (Tables 1-2)
- 4,170 (MIMIC-III) and 5,152 (UPHS) ICU notes were used to measure lexicographic coverage (Figure 1) and predict in-patient mortality (Figure 2)
- The first physician encounter note charted within 24 hours of hospital admission was used to predict in-patient mortality
- Text of a note was converted into sentence vectors by taking the centroid of a word vector matrix, where every row in the matrix is a vector representation of a word in that sentence
- Prediction models were evaluated using the Brier Score with a 95% confidence interval

Results

Training algorithm	Text Corpus	Vector Dimension
	MIMIC-III	
Word2Vec		100
	PMC Open Access Subset- All manuscripts	
		300
FastText	PMC Open Access Subset- Case reports only	
		600
	University of Pennsylvania Health System	
GLoVE		1200
	Wikipedia	

Table 1. Summary of model types, text corpora, and vector dimensions used for training word embeddings.

Corpus	Corpus Documents	Total Tokens
MIMIC-III	27,449	49,590,835
PMC Open Access Subset- All manuscripts	220,453	148,089,760
PMC Open Access Subset- Case reports only	628,404	1,848,856,520
University of Pennsylvania Health System	4,555,827	2,542,552,916
Wikipedia	14,828,230	10,917,117,453

Table 2. Summary of training corpora statistics.

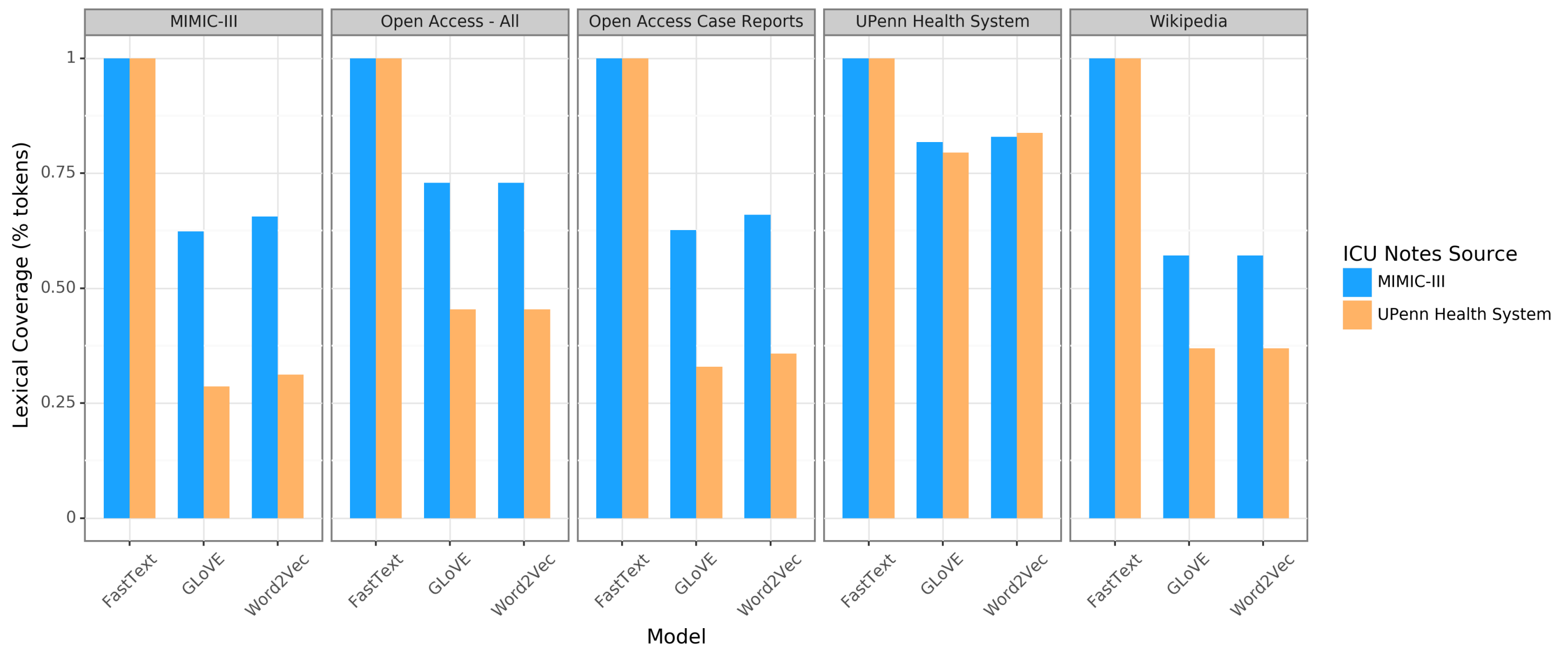

Figure 1. Fraction of unique tokens in ICU notes that contain vector representations in a given embedding model.

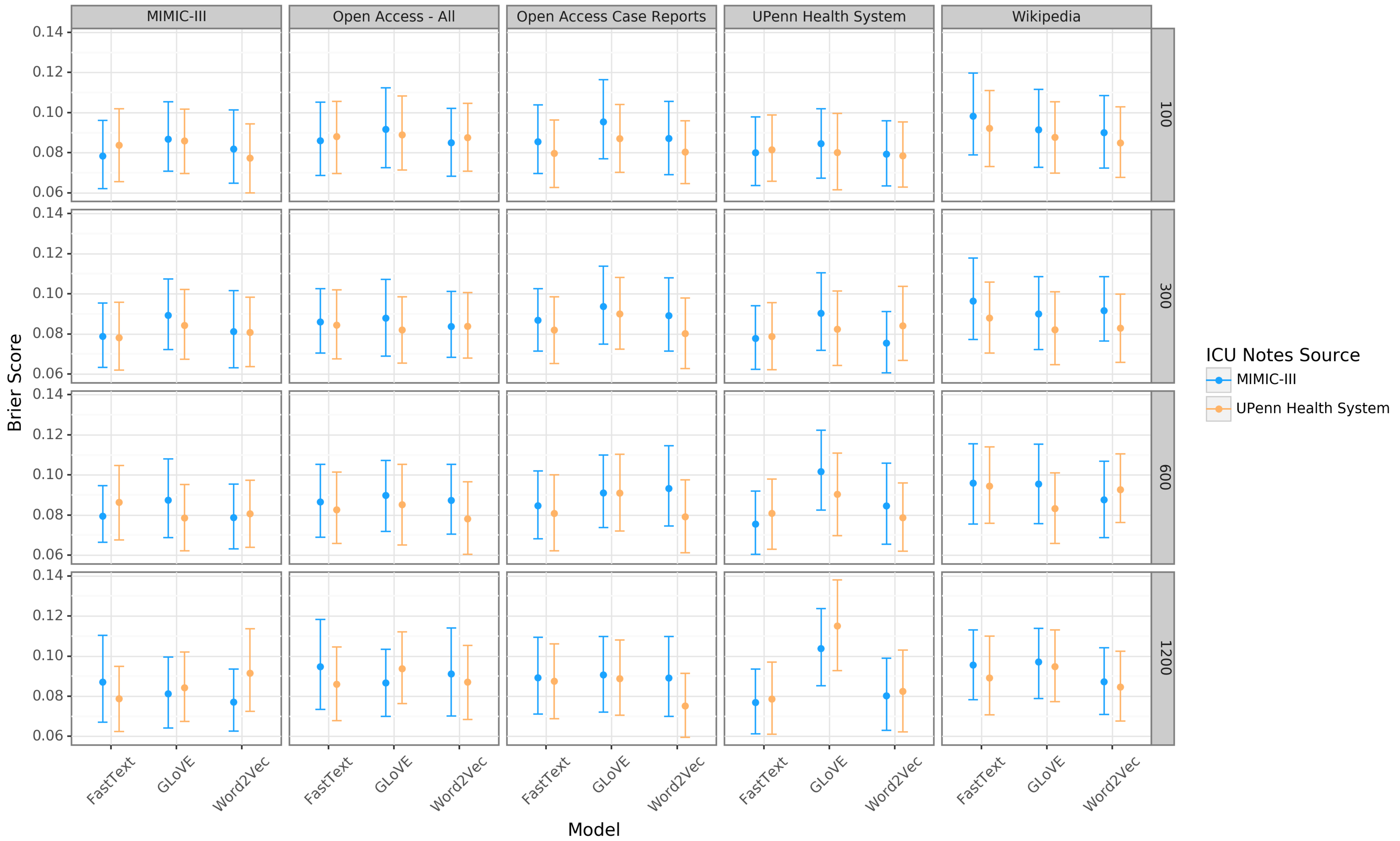

Figure 2. Performance of a mortality prediction model using the text of the first physician encounter note for each hospitalization.

Conclusions

- FastText embedding models obtain 100% lexicographic coverage but do not perform better than Word2Vec or GLoVE embedding models in prediction of in-patient mortality
- The smaller Open Access Case Reports embeddings performed as well as embeddings trained on much larger corpora in predicting in-patient mortality
- Open Access Case Reports embeddings are available for download at https://github.com/weissman-lab/clinical_embeddings

Limitations

- It is unknown whether the comparable performance seen for embeddings trained on case reports will be true for new concept embedding representation methods
- While we utilized a centroid based method for converting text into numeric input for neural network it is not clear this is the best method

Albert Einstein College of Medicine
Medical Scientist Training Program

pair
Palliative and Advanced Illness Research Center

Perelman
SCHOOL OF MEDICINE
UNIVERSITY of PENNSYLVANIA

Contact

✉ zachary.flamholz@einsteinmed.org
@zflam94