Reading You Like a Book: Aggregated Medical Code Embeddings for Predicting Hospital Readmissions

Sacha Davis[1], Jin Zhang[2], Ilbin Lee[2], Mostafa Rezaei[4], Russell Greiner[1 5], Finlay A. McAlister[3] and Raj Padwal[3]
[1]Department of Computing Science, [2]Alberta School of Business, [3]Medicine and Dentistry, [4]ECSP Business School, [5]Alberta Machine Intelligence Institute

Problem Motivation

Common types of Electronic Health Record data:



Reasons Electronic Health Record data is collected:



Tabular data used to predict clinical outcomes is:

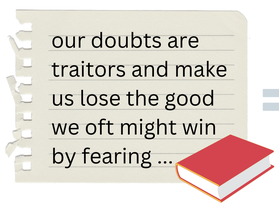
- Of poor quality (many missing values) OR
- Painstakingly hand-engineered

Hospitals are currently too overloaded to overhaul their data collection practices.



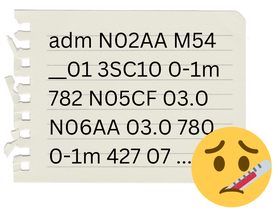
How can we incorporate unstructured historical data into patient representations for ML?

Natural Language Document



|Vocab| = ~20000+

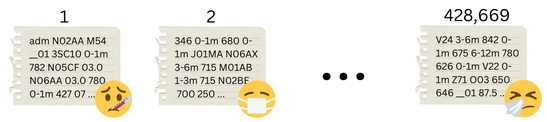
Patient History "Document"



|Vocab| = 3664

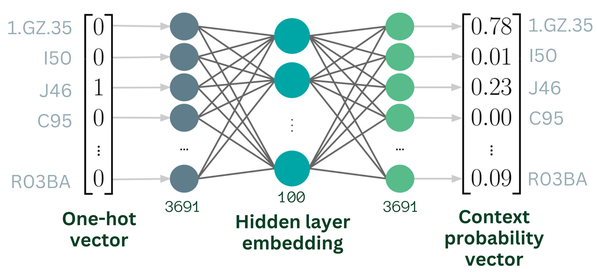
Medical Code Embeddings

Population-wide medical usage from Alberta Health Services from Jan 01 2012 - Dec 31 2014:



From training set, calculate **Context Probability Vector** for each code **c**. Each element is the probability that a code appears in the context of **c**.

Train Word2Vec skip-gram model to predict CPVs:



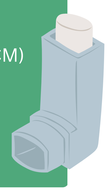
Extract hidden layer embeddings:

	1	2	3	4	...	100
1.GZ.35	0.56	-6.09	1.01	4.33	...	0.71
I50	1.15	-0.78	0.51	0.02	...	-1.56
ADM	0.04	3.41	-1.89	0.89	...	0.43
C95	-0.99	-0.34	0.08	-2.30	...	1.91
...	...	...	...	...	...	...
R03BA	2.9	-0.01	3.79	0.56	...	-0.22

Codes in similar contexts receive similar vectors:

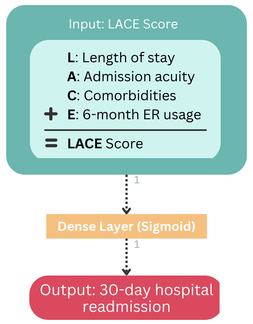
Most cosine similar codes to **J46**: Asthma (ICD-10-CM):

- **493**: Asthma (ICD-9)
- **1.GZ.35**: Inhalation pharmacotherapy (CCI)
- **J98**: Unspecified respiratory disorders (ICD-10-CM)
- **R03AC**: Drugs for obstructive airway disease: Selective beta-2-adrenoreceptor agonists (ATC)
- **R03BA**: Drugs for obstructive airway diseases: Glucocorticoids (ATC)

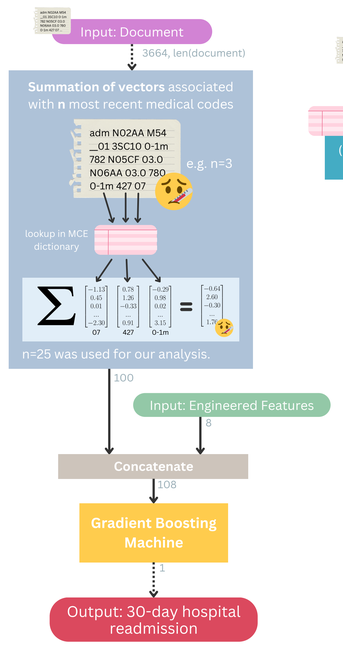


Aggregating MCEs and Predicting Readmission

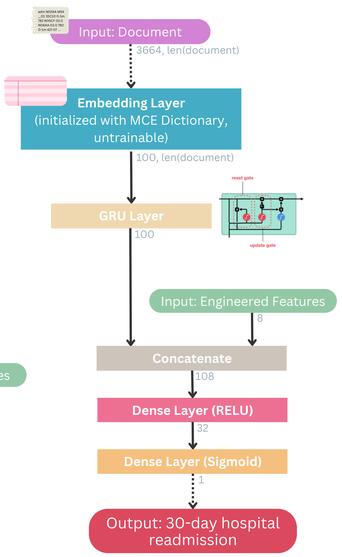
LACE Score (Baseline)



MCE Summation + GBM

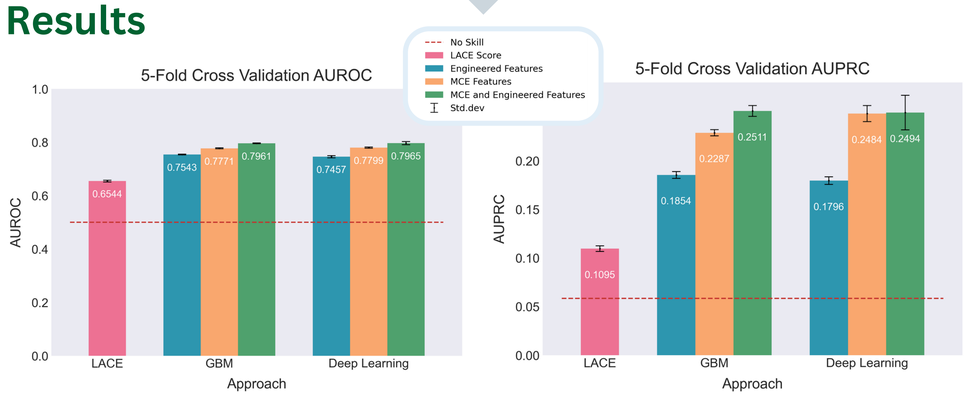


MCE-Initialized GRU-RNN



Hospitals often rely on heuristics to predict clinical events. Using the **LACE** index is common practice that assigns patients a score from 0-19, and is used to rank patients by risk of readmission.

Results



MCE features improve upon the 30-day all cause readmission task.