Identifying Transitional High Cost Users from Unstructured Patient Profiles Written by Primary Care Physicians

Haoran Zhang[1,2], Elisa Candido[3], Andrew S. Wilton[3], Raquel Duchen[3], Liisa Jaakkimainen[3], Walter Wodchis[1,3,4], Quaid Morris[1,2,5]

VECTOR INSTITUTE

[1]University of Toronto [2]Vector Institute [3]ICES [4]Trillium Health Partners [5]Memorial Sloan Kettering Cancer Center

UNIVERSITY OF TORONTO

Motivation

- One strategy to reduce growing healthcare costs is to target the High-Cost Users (HCUs).
- In Ontario, the top 5% of patients consume up to 66% of hospital and homecare costs [4].
- About half of HCUs are recurrent year-to-year.
- Patients often become recurrent HCUs through an adverse health event, resulting in a transition to a more frail health state that is difficult to reverse.
- To properly target HCUs, should look to prevent this transition in the first place.
- Family physicians are in the best position to prevent this transition.

In this study, we build predictive models that identify patients at risk of becoming HCUs using only data available to family physicians.

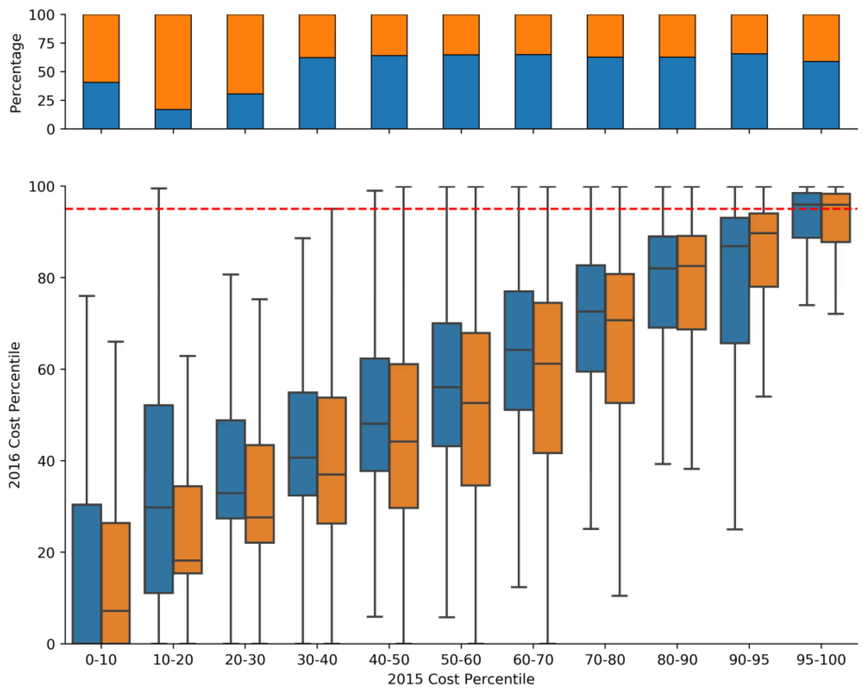
EMR Data

- Use the Electronic Medical Record Primary Care (EMRPC) database, housed at ICES in Toronto, Ontario.
- Contains EMR records for ∼600,000 patients in Ontario.
- Use the Cumulative Patient Profile (CPP), which contains the following text fields (with sample values):

Field Name	Sample Content
Allergies	*Demerol 50 mg Tablet ->vomiting, delerious*
Problem list	*non smoker; alcohol consumption – none; PAF; rt shoulder pain; 2008-dysmenorrhea-*
Risk factors	*never smoked; f/u with special views & U/S lt breast; Pap: Feb 12 N; FHN*
Personal traits	*Teaches grade 6-7; Arts/theatre; diet: low fruit and veg, fish; exercise – walking 3hrs/ wkly*
Family history	*mother a&w; Farther – d 86 – Brain Ca; f-dm2; hypertrophic cardiomyopathy – pat aunt*
Past medical history	*PAF; T&A; CVS; GI; Spinal Curvature – since childhood; g4P4, 1st 2 Csxn for FTP, 2nd 2 repeat sections*

Cost Data

- Created a cohort of 277,173 patients.
- Calculated person-level healthcare costs based on billing codes and other service utilizations.
- Goal: predict whether a patient will be a top 5% HCU in 2016 from 2015 data.
- Observation: Future cost percentile is highly correlated with current cost percentile (orange = men, blue = women).



Word Embeddings

- Tested word embeddings from the word2vec skip-gram model pretrained on the following four corpora:

Training Data	Training Tokens (billions)	Dimension	Vocab Size	Cased
PubMed [2]	2.7	200	2,231,686	Yes
PubMed+PMC+Wiki [5]	>5.5	300	5,443,656	Yes
MIMIC-III	0.6	300	420,786	No
EMRPC	2.3	300	723,458	No

Term Similarities

- Obtained closest words by cosine distance to common clinical terms.
- Embeddings originating from clinical notes captured common misspellings and abbreviations, while embeddings from well-edited text identified related concepts.

	PubMed		PubMed+PMC+Wiki		MIMIC		EMRPC	
throat	sore	0.79	throats	0.74	nose	0.79	thoat	0.85
	throats	0.77	runny	0.70	ears	0.63	thraot	0.82
	pharyngitis	0.75	pus/surface	0.69	thin	0.57	thorat	0.79
	Throat	0.74	nose	0.69	oropharynx	0.56	troat	0.78
	pharyngotonsillar	0.73	sore	0.68	normocephalic	0.56	throat-	0.73
diabetes	mellitus	0.95	T2DM	0.85	iddm	0.76	dm	0.76
	T2DM	0.83	T1DM	0.82	mellitus	0.76	dm2	0.69
	Diabetes	0.82	prediabetes	0.80	dm	0.73	diabetic	0.67
	diabetic	0.82	mellitus	0.80	diabetic	0.73	t2dm	0.63
	non-insulin-dependent	0.80	pre-diabetes	0.78	asthma	0.66	diabtes	0.61
aspirin	Aspirin	0.87	clopidogrel	0.85	plavix	0.72	asprin	0.80
	acetylsalicylic	0.85	ticlopidine	0.84	asa	0.69	asa	0.76
	clopidogrel	0.79	Aspirin	0.80	statin	0.62	aspririn	0.69
	antiplatelet	0.75	clopidegrol	0.79	lisinopril	0.60	81mg	0.68
	ER-DP	0.75	warfarin	0.78	atorvastatin	0.59	plavix	0.67

- For modelling, each field in the CPP was encoded separately using an aggregated word embedding method (referred to as *EmbEncode*) [1, 3].

Overall Model Performance

- Added age and sex as features
- Experimented with adding current cost percentile (even though family physicians do not have access to this field).
- Logistic regression with Bayesian hyperparameter search

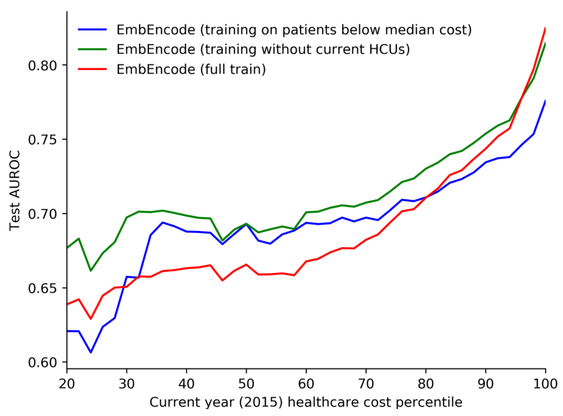
	% ROC (± 95% CI): CPP only			% ROC (± 95% CI): CPP + current cost percentile		
Embedding	BoW	EmbEncode	BoW+EmbEncode	BoW	EmbEncode	BoW+EmbEncode
PubMed	81.85 (0.36)	81.96 (0.36)	82.23 (0.35)	89.74 (0.27)	90.13 (0.27)	89.74 (0.28)
PubMed + PMC + Wiki		82.05 (0.35)	82.23 (0.35)		**90.20 (0.27)**	89.77 (0.28)
MIMIC		81.90 (0.36)	82.27 (0.35)		90.08 (0.27)	89.76 (0.27)
EMRPC		**82.48 (0.35)**	**82.45 (0.35)**		90.19 (0.27)	**89.81 (0.28)**

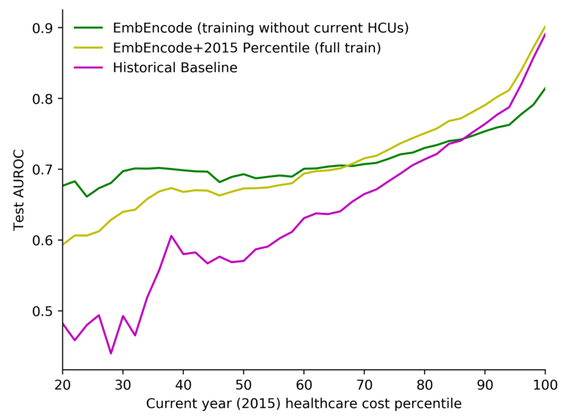
Takeaways:

- The current cost percentile is a highly informative feature overall. In fact, using this feature on its own achieves 89.12% ROC.
- Embeddings trained on the same dataset as the downstream task (EMRPC) perform much better than embeddings from other domains.
- The EmbEncode model outperforms the baseline Bag of Words (BoW) model.

Performance on Potential Transitional HCUs

- Experimented with exclusion of current year HCUs from training set.
- Experimented with evaluating ROC performance on subsets of test set patients thresholding by current year percentile.





Takeaways:

- Removing current HCUs from the training cohort appears to be an effective way of improving model performance on patients with lower current year healthcare costs.
- Inclusion of the current year cost percentile as a feature decreases model performance on potential transitional users.

References

[1] W. Boag et al. What's in a note? unpacking predictive value in clinical note representations. *AMIA Jt Summits Transl Sci Proc*, 2018:26, 2018.
[2] B. Chiu et al. How to Train good Word Embeddings for Biomedical NLP. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, 2016.
[3] S. Dubois et al. Learning Effective Representations from Clinical Notes. *arXiv:1705.07025*, 2017.
[4] E. Homenauth et al. Examination of High-Cost Patients in Ontario. *International Journal of Population Data Science*, 3(3):359, 2018.
[5] S. Pyysalo et al. Distributional Semantics Resources for Biomedical Text Processing. *Proceedings of LBM*, pages 39–44, 2013.