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

Haoran Zhang [i,ii,iii,iv], Elisa Candido[iv], Andrew S. Wilton[iv], Raquel Duchen[iv], Liisa Jaakkimainen[iv], Walter Wodchis[iv, v,vi], Quaid Morris[i, ii, iii vii,viii]

Identification and subsequent intervention of patients at risk of becoming High Cost Users (HCUs) presents the opportunity to improve outcomes while also providing significant savings for the healthcare system. In this paper, the 2016 HCU status of patients was predicted using free-form text data from the 2015 cumulative patient profiles within the electronic medical records of family care practices in Ontario. These unstructured notes make substantial use of domain-specific spellings and abbreviations; we show that word embeddings derived from the same context provide more informative features than pre-trained ones based on Wikipedia, MIMIC, and Pubmed. We further demonstrate that a model using features derived from aggregated word embeddings (EmbEncode) provides a significant performance improvement over the bag-of-words representation (82.48±0.35% versus 81.85±0.36% held-out AUROC, $p = 3.2 \times 10^{-4}$), using far fewer input features (5,492 versus 214,750) and fewer non-zero coefficients (1,177 versus 4,284). The future HCUs of greatest interest are the transitional ones who are not already HCUs, because they provide the greatest scope for interventions. Predicting these new HCU is challenging because most HCUs recur. We show that removing recurrent HCUs from the training set improves the ability of EmbEncode to predict new HCUs, while only slightly decreasing its ability to predict recurrent ones.

Keywords: Electronic Medical Records, High Cost Users, Natural Language Processing, Deep Learning

1. Introduction

Many healthcare systems face significant challenges in caring for their aging populations, the solutions to which may lie in identifying patients at risk of undergoing serious adverse health events. One strategy to reducing healthcare costs lies in targeting the so-called high cost users (HCUs): the small group (e.g., 5% of the population) which consumes the majority of the healthcare resources.[8,19,36,47] For example, in Ontario it has been reported that the top 5% of patients consume anywhere from 61%[36] to 66%[17] to 84%[29] of hospital and homecare costs. Indeed, there are ongoing programs across several Canadian provinces focused on addressing

[i]Department of Computer Science, University of Toronto
[ii]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada
[iii]Corresponding authors: haoran@cs.toronto.edu, quaid.morris@utoronto.ca
[iv]ICES, Toronto, Ontario, Canada
[v]Institute of Health Policy, Management, and Evaluation, University of Toronto
[vi]Institute for Better Health, Trillium Health Partners, Mississauga, Ontario, Canada
[vii]Terrence Donnelly Center for Cellular and Biomolecular Research, University of Toronto
[viii]Department of Molecular Genetics, University of Toronto

avoidable costs by targeting HCUs.[6] Patients often become recurrent HCUs due some serious adverse health event (such as a fall or a catastrophic worsening of an existing chronic condition), resulting in a transition to a more frail health state[19] that is difficult to reverse.[45] The existence of a persistent HCU cohort is well-known.[33] One Canadian study showed that about 45% of HCUs in Ontario recur year-to-year, and another study in the US showed that HCUs were nearly as likely to remain HCUs five years later as one year later.[16] In many cases, the initial transition to HCU might have been avoidable with better tailored, preventative care.[1,2] However, to provide this care, these at-risk patients must first be identified.

There have been several previous studies that use predictive modelling to directly identify potential HCUs;[1,4,12,13,28,30,41] however, most of this work is not relevant to finding new HCU patients before their transition. Specifically, because they use data collected in a hospital setting, many of their patients have already experienced the transitional event. Here, we focus on identifying at-risk patients outside of hospitals, based on data collected by their primary care (i.e. family) physician.[39] By identifying patients in this setting, we are more likely to catch patients at risk of becoming HCUs before their transitional event(s). The most readily available source of primary care data is their patient notes. Unlike much of the data collected in a hospital setting, and used in most of these other studies, these notes are rarely structured.

Here we describe a predictive tool that identifies patients at risk of becoming HCUs using only data readily available to family physicians – unstructured patient profiles stored in their Electronic Medical Record (EMR). Family physicians are the first and most frequent point of contact for the patient,[31,38] and thus have the most up-to-date information about the patient's health. Very few studies have used primarily free-form text within the EMR for the purpose of predicting high cost users. One exception is Frost et al,[13] which used logistic regression on a bag-of-words (BoW) representation of the patient profiles. However, unlike well-edited documents usually seen in text mining, the patient profiles in our EMRs contain numerous acronyms, short forms, misspellings, and potentially important technical terms that appear infrequently.[13,15,42] These complexities limited the effectiveness of previous models.[13] In this paper, we overcome this challenge using word embeddings, a dense continuous-value vector that depends on the other words that they frequently appear with.[20,27,32] These embedding are trained using deep learning methods, and have been shown to be useful for representing clinical text.[3,7,9,21–24]

Here we examine the value of embeddings for this challenging classification setting. The patient profiles that constitute our inputs are brief and use non-standard language. The labels for our patients are, by definion, highly imbalanced: HCUs are only 5% of patients and the new HCUs, which are of greatest interest, are only a small fraction of this group. We demonstrate that in this setting, word embeddings can be used to derive a relatively small number of highly informative features, thereby permitting accurate classification of patients most at risk of becoming HCUs.

In section 2, we describe the procedure used to create a study cohort and training sets from the EMR and cost data. In section 3, we describe the embedding methods and models used to predict HCUs. In sections 4 and 5, we show the results obtained and discuss their significance.

2. Data

2.1. *EMRPC*

The Electronic Medical Record Primary Care (EMRPC) database, housed at ICES in Toronto, Ontario,[43] contains unstructured text data for approximately 600,000 patients from family practices that have opted in. For this paper, the cumulative patient profile (CPP) section of EMRPC was used for modelling. The CPP consists of six free-form text fields: allergies, problem list, risk factors, personal traits, family history, and past medical history. These fields are updated by the family physician and provide a summary of their cumulative observations regarding the patient. Table 1 shows sample, anonymized values for each field of the CPP. Similar to this example, these text fields contain many abbreviations, typos, and specialized vocabulary.

Table 1. Sample CPP fields (modified to preserve patient confidentiality)

Field Name	**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*

2.2. *Total Healthcare Costs*

Ontario has a universal, publicly funded, healthcare system. To calculate the total health system cost for an individual over a certain period, the GETCOST macro, developed by ICES, was used. GETCOST* is based on a previously described methodology for computing total healthcare costs from administrative data.[46]

To create the cohort from EMRPC, all individuals whose EMR data were collected and entered into EMRPC between 1 April 2015 and 31 March 2016 were selected. The date of data collection is known as the "load date". Individuals younger than 18 or older than 105 years of age and those who had been on the EMR system for less than one year as of the load date were excluded. Using the GETCOST macro, the total healthcare costs, along with their cost percentile rankings among all Ontarians, were calculated for each individual for 365 days after their profile load date (i.e. the next year), and for 365 days prior to the load date (i.e. the current year). Individuals with missing costs in either year (e.g. due to death) were excluded.

*GETCOST computes person-level costs based on billing codes and other service utilizations paid for by the Ontario government.

The final cohort size was 277,173 patients. Text from each of the six CPP fields were converted to lower case and the vocabulary of tokens for each field was determined independently. Stop words and tokens starting with a digit were removed. We define HCUs as individuals in the top 5% of total healthcare costs. Figure 1 compares a patient's cost percentile for the next year (i.e., 365 days after the load date) versus the current year (i.e. 365 days prior to the load date). Consistent with previous reports,[45] these two costs are highly correlated ($R^2 = 0.688, \rho = 0.686$). Also, as per previous reports,[11,37,45] we found a high correlation between the cost percentile and age ($R^2 = 0.465, \rho = 0.458$). As such, we included both age and sex as features when modelling.

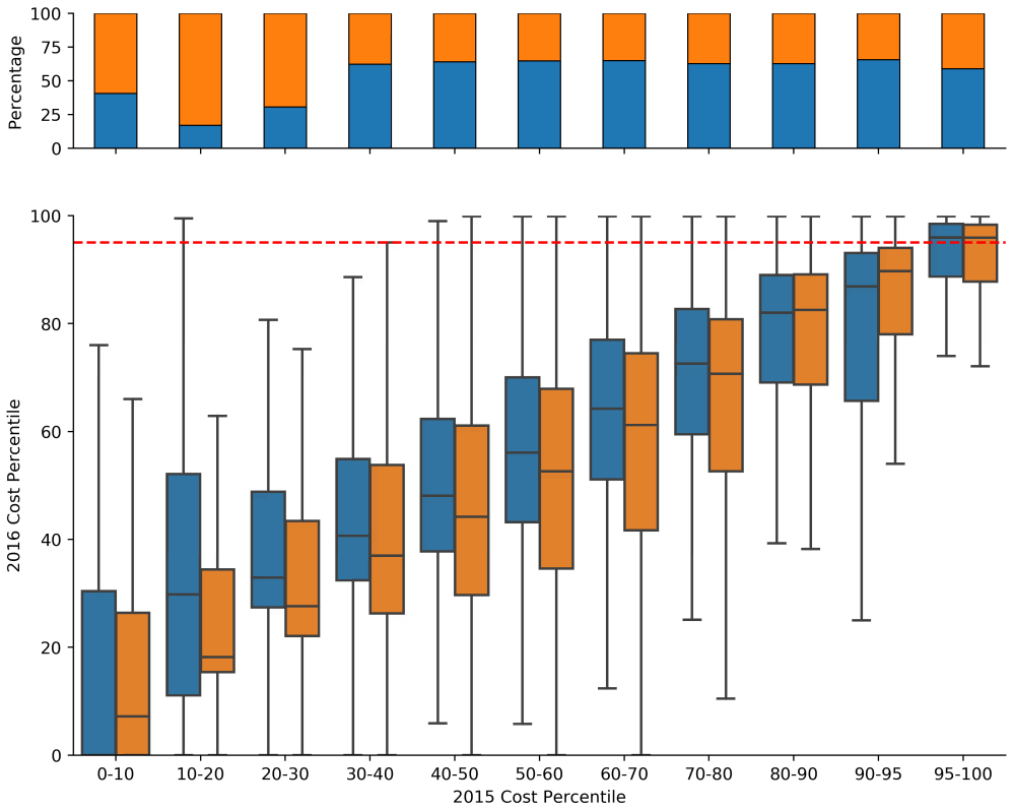

Figure 1. Box plots show distribution of next year (2016) cost percentile for cohort males (orange) and females (blue) in the current year (2015) cost percentile range shown on the x-axis. The stacked bars show the sex distribution within each current year cost percentile range.

After data processing, the cohort was randomly split into training and test sets, stratifying by their next year HCU status. The training set consisted of 80% of the total cohort (221,738 records), and the test set consisted of 20% of the total cohort (55,435 records). All model parameters and hyperparameters were set using the training set, all model performance estimates reported here are based solely on the test set.

2.3. *Encoding of Ordinal Variables*

To encode the ordinal variables age and current year percentile as features, first, for each variable, we choose a set of thresholds $t = \{t_1, t_2, ..., t_N\}$ where $t_i < t_j$ if $i < j$, and N is the

number of encoded binary features. Then, if z is the value to be encoded, we set $x_i = 1$ if $z \geq t_i$, otherwise $x_i = 0$. After a logistic regression model is fit, this allows us to visualize the weights $\{w_1, w_2, ..., w_N\}$ by plotting the cumulative weight $\sum_{j=1}^{i}(w_j)$ versus t_i.

To encode age, we used $t = \{19, 20, ..., 103\}$, and to encode the current year percentile, we used $t = \{6, 7, ..., 95, 95.1, ..., 99.9\}$.

2.4. *Word Embeddings*

In language prediction tasks, like ours, with a small set of positively labelled examples, externally defined word embeddings are often used.[14,26] However, as the notes in our study used numerous abbreviations unique to this context, we reasoned this strategy would be less effective than word embeddings specific to our corpus. To assess the impact of source of word embeddings, we tested four different sets of word embeddings, two were pre-defined by others,[5,35] and two were created for this study. All four embedding sets were originated using the word2vec skip-gram model.[27] Table 2 shows a summary of the embeddings used. As seen in Table 3, embeddings defined using clinical notes (i.e. EMRPC and MIMIC) were able to capture common misspellings and abbreviations of terms, while more generalized embedding sets derived from well-edited text identified similar or related concepts. Note that the EMPRC embedding used not only the CPP used by our classifier, but also the less structured "progress notes" not considered by our classifier.

Table 2. Summary of the embedding sets used: the number of tokens used for training, the number of dimensions, the number of unique tokens, and whether or not the vocabulary is cased. The PubMed and PubMed+PMC+Wiki embeddings were taken from external sources, while the MIMIC and EMRPC embeddings were trained for this study.

Training Data	**Training Tokens (billions)**	**Dimension**	**Vocab Size**	**Cased**
PubMed[5]	2.7	200	2,231,686	Yes
PubMed+PMC+Wiki[35]	>5.5	300	5,443,656	Yes
MIMIC[18]	0.6	300	420,786	No
EMRPC	2.3	300	723,458	No

3. Methods

For all models, logistic regression with L1 regularization was used as the classifier. The regularization parameter was tuned with a Bayesian optimization for 25 iterations using Scikit-optimize,[25] with stratified 5-fold cross validation on the training cohort and AUROC as the metric. The three main models we compared correspond to three different feature sets which are defined below. The sets are: Bag of Words (BoW), EmbEncode, and EmbEncode appended to BoW (BoW+EmbEncode).

3.1. *Bag of Words*

After preprocessing and Porter stemming,[34] unigrams and bigrams appearing in at least five training set documents were included in the vocabulary of the bag of words model. Separate

Table 3. Top 5 closest words based on cosine distance to some common clinical terms for each of the embeddings

	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
cancer	cancers	0.88	cancers	0.86	metastatic	0.83	ca	0.77
	breast	0.80	caner	0.82	carcinoma	0.78	colon	0.75
	carcinoma	0.79	CRC	0.8	mets	0.78	cancer-	0.69
	colorectal	0.78	PCa	0.79	melanoma	0.74	colorectal	0.64
	adenocarcinoma	0.77	cancer.4	0.75	chemo	0.71	cancern	0.62
metastasis	metastases	0.88	metastases	0.89	metastases	0.89	metastases	0.87
	metastatic	0.82	micrometastasis	0.82	metastatic	0.75	metastatic	0.79
	metastasized	0.78	metastatis	0.80	carcinoma	0.72	mets	0.78
	tumor	0.78	metastastic	0.80	mets	0.71	tumor	0.70
	metastastic	0.77	metastatic	0.79	neoplasm	0.69	tumour	0.69
arthritis	rheumatoid	0.92	arthritis(RA)	0.79	gout	0.80	oa	0.78
	polyarthritis	0.83	polyarthritis	0.78	osteoarthritis	0.76	osteoarthritis	0.77
	arthritides	0.82	arthritides	0.77	rheumatoid	0.73	arthrits	0.76
	ankylosing	0.80	PsA	0.76	oa	0.72	arthrtis	0.76
	spondylitis	0.80	arthritis-like	0.75	neuropathy	0.66	arthirits	0.74
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
dialysis	hemodialysis	0.87	haemodialysis	0.80	hd	0.82	hemodialysis	0.72
	CAPD	0.84	hemodialysis	0.79	hemodialysis	0.73	dyalisis	0.66
	Dialysis	0.82	CAPD	0.77	quinton	0.70	dialysis-	0.64
	haemodialysis	0.82	CRRT	0.73	tunneled	0.66	diaylsis	0.64
	dialytic	0.80	dialytic	0.73	permacath	0.63	dialyisis	0.63

vocabularies were defined for each of the six fields in the CPP; and separate word counts were collected for each of these fields. We applied TF-IDF weighting to these appended word counts, to define the final input vector. Age and sex were also added as features, with age encoded as described in section 2.3. Features were each scaled to unit variance. The dimension of the input feature vector for this model was 214,750.

3.2. *EmbEncode*

In the EmbEncode model, each field of the CPP is encoded by a vector of length equal to three times the embedding dimension. The first third of the vector corresponds to the elementwise max of the embedding vectors of the tokens within the document, the second corresponds to the elementwise minimum, and the last third corresponds to the elementwise mean. Four instances of the EmbEncode model were created using the four pre-trained word2vec embeddings defined in Section 2.4. Aggregated word embedding features have proved useful in other clinical prediction tasks.[3,10] Age and sex were added, as per the last section, and the

features were scaled to unit variance. Features derived from word embeddings better capture word semantics, as words with similar semantic meaning have similar embeddings. Aggregating word embeddings using the EmbEncode features permit us to represent documents of different sizes with a fixed length feature vector.

3.3. *Historical Baseline*

As seen in Figure 1, the cost percentile of a person in the current year is highly correlated with their cost percentile in the next year. The historical baseline is a classifier which simply has the current year cost percentile as its only feature. It assigns a discriminant value to each patient equal to their cost percentile in the current year. For example, if a person was in the 84th cost percentile in the current year, their discriminant value for being in the top 5% HCU in the next year would be 0.84. Patients are ordered by this discriminant value to compute ROC areas. For some of the models, we also encode current year cost percentile as a feature.

3.4. *Varying the Training Set*

We explore the exclusion of certain individuals from the training set based on their current year cost percentile. By removing patients with high cost percentile in the current year, we hope that the model will better predict transitional users, rather than focusing on recurrent HCUs. However, it is not clear that removing recurrent HCUs would improve classifier performance because excluding current HCUs greatly reduces the number of positive examples available to train the model. As such, we explore the following three training sets: 1) All patients ($n = 221,738$, with $16,019$ becoming or remaining HCUs in the next year), 2) Patients with current year cost percentile less than 95% ($n = 206,935$, with $7,675$ becoming HCUs in the next year), 3) Patients with current year cost percentile less than or equal to 50% ($n = 92,833$, with 684 becoming HCUs in the next year). Note that removing recurrent HCUs reduces the number of positive examples available to train our classifier by more than 50%.

3.5. *Varying the Evaluation Set*

In order to effectively evaluate model performance only on potentially transitional HCUs, we also explore adjusting patient inclusion in the test set. In particular, model performance will be evaluated on different subsets of the test set, created based on the patients' current year percentiles. We use subsets of patients with current year cost percentile less than or equal to k, where $k = 20, 22, 24..., 100$.

4. Results

First, using the entire training set, we ran the Bayesian optimization for the three models on each of the four embeddings, selecting the set of hyperparameters with the best 5-fold ROC performance for each (embedding, model) combination. One additional variant of each model was trained, adding on the current year cost percentile as a feature, while maintaining the optimal hyperparameters from the Bayesian search. Each model was evaluated on the held-out test set, and the ROCs of the models are shown in Table 4. 95% Confidence intervals

(shown in parentheses) were calculated using the DeLong test.[40] Without using the current year cost percentile, the best performing model is the EmbEncode with EMRPC embeddings. Despite having 50-fold fewer features (219,034 versus 4,284), EmbEncode performs better than BoW+EmbEncode on the EMRPC embeddings. This suggests that despite the regularization and hyperparameter tuning, there is still some overfitting. Note that after L1 regularization, the EmbEncode model contains 1,177 features with non-zero weights, while adding in the BoW features increases it to 4,284. Note that for embeddings other than EMRPC, adding the BoW features increased performance slightly, potentially because the embeddings were not well matched to the classification problem.

Table 4. Test set % ROC of each embedding set for the best hyperparameter combination obtained from the Bayesian search. Also includes the test set ROC for the best hyperparameters plus adding the current year cost percentile. All models were trained on the entire training set.

	% 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)**

From Table 4, it is clear that the current year percentile is a highly predictive variable. In fact, the historical baseline has a performance of 89.12±0.29% ROC, far exceeding the performance of any of the models without current year cost percentile as a feature. However, as suggested in Figure 1, current year percentile may not be a useful feature for patients with lower percentiles.

Next, to examine the impact of current year cost, and of adjusting the training set to focus on non-recurrent HCUs, we evaluated models trained with different subsets of the training data on different subsets of the testing data. As shown in Figure 2, removing recurrent high cost users from the training set appears to be an effective method of increasing model performance for users with initially low healthcare costs. Note, in particular, that the 95% cutoff model performs better than the 50% cutoff model even on those patients below the 50% cutoff that the latter was targeted toward. Selecting the best model from this figure (EmbEncode with 95% training cutoff) and comparing it with other models in Figure 3, it can be seen that it outperforms a model with current year percentile added until up to 70% current year percentile (in other words, for the majority of the patient population). Furthermore, the historical baseline performs much worse compared to other models on users with lower initial costs.

Figure 4 shows the cumulative weights (as outlined in Section 2.3) of the age variable for an EmbEncode model based on EMRPC embeddings. Figure 5 shows the cumulative weights for the current year percentile variable for an *EmbEncode+current year percentile* model based on EMRPC embeddings. As expected, both plots show a generally increasing trend. However, the contribution of age to the log-probability seems to diminish with increasing age past 80 years old. Conversely, the contribution of the current year cost percentile seems to grow rapidly for the last 5-10 percentiles. In addition, note that the cumulative coefficient for the current

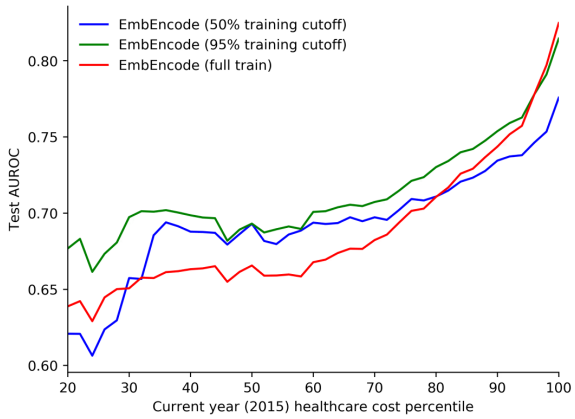

Figure 2. ROC Performance of EmbEncode models with EMRPC embeddings trained on different subsets of the training cohort evaluated on different subsets of the test cohort based on their current year cost percentile

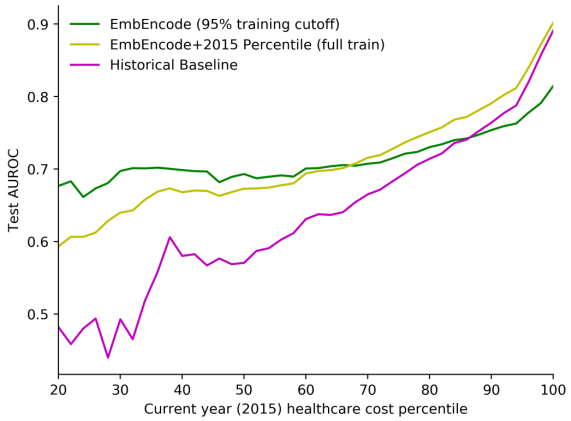

Figure 3. ROC Performance of various models evaluated on different subsets of the test cohort based on their current year cost percentile

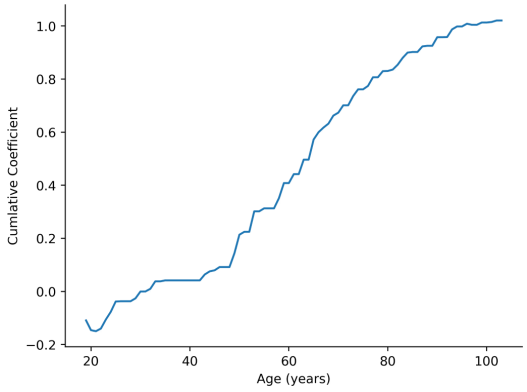

Figure 4. Cumulative coefficients for the age variable

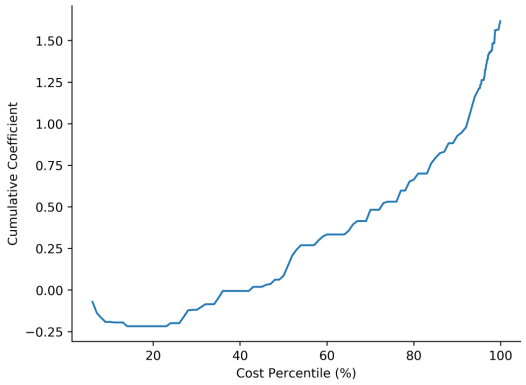

Figure 5. Cumulative coefficients for the current year (2015) cost percentile variable

year percentile is close to zero until about 50%, further indicating that current year healthcare costs is not an useful indicator of next year spending for lower-cost patients.

5. Discussion

Choice of Word Embeddings From Table 4, it can be observed that the EMRPC embeddings perform much better than the other embeddings. By training the embeddings on the same dataset as the downstream task, the embeddings were able to encode misspellings and abbreviations that appear in the input text with similar vectors as the original concept, leading to better downstream performance.[44] The poor transferability of embeddings trained on journal articles to clinical text is also documented in the literature.[10]

Performance of Historical Baseline As anticipated, a patient's healthcare usage in the current year is a good predictor of their HCU status in the next year, but only if the patient

is already a HCU. As such, this information, which is difficult for primary care providers to access, has little value for patients not already in the top half of healthcare users. Indeed, when the current year percentile is added to the EmbEncode model, the performance of the model actually deteriorates for patients with lower levels of healthcare usage.

Varying the Training Set Removing patients with high costs in the current year from the training cohort appears to be an effective way of improving model performance on patients with lower current year healthcare costs. In particular, training on patients with less than 95% current year cost percentile results in an improvement over selecting only patients with less than 50% current year percentile. However, both models noticably out-perform training on the entire cohort. This indicates that the risk of overfitting on recurrent HCUs outweighs the extra data available from using all the patients for training. Further, both models out-perform the EmbEncode model with the current year percentile included (the model with nearly the highest overall ROC), for the majority of patients.

Performance of EmbEncode Using EMRPC embeddings, the EmbEncode model performs competitively with the BoW+EmbEncode model using only a quarter of the features. For other embedding sets, adding aggregated embedding features to the bag-of-words model improved the model performance for all embedding sets tested. Many of these embedding features have large coefficient magnitudes, indicating that they are important variables for the models. It is clear that aggregated word embeddings can efficiently capture useful semantic information.

Limitations and Future Work In this study, we focused on patient data readily available at the point-of-care to primary care physicians in Ontario and have established a strong baseline for predictive performance. However, the CPP is simply a summary of the entire patient record available to the primary care physician; incorporating temporal information through progress and consult notes may improve performance but would require more complex models. Other administrative data, generally not available to primary care providers in Ontario, such as emergency room visits, hospitalizations, and other healthcare utilization statistics may also improve the model performance. We have also not considered word ordering, except for bigrams in the BoW model. In some cases, these features can be helpful, but would also require more complex and less interpretable models. Finally, our findings focus on unstructured cumulative patient profiles, so the methods presented here would not generalize to other EMR systems that do not contain them.

Acknowledgments

Quaid Morris holds a Canada CIFAR AI chair. The datasets used in this project were provided by ICES. This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsements by ICES or the Ontario MOHLTC should be inferred. Parts of this material are based on data and/or information compiled and provided by CIHI. However, the analyses, conclusions, opinions and statements expressed in the material are those of the authors, and not necessarily those of CIHI.

References

1. D. Anderson and M. Bjarnadóttir. When is an ounce of prevention worth a pound of cure? Identifying high-risk candidates for case management. *IIE Transactions on Healthcare Systems Engineering*, 6(1):22–32, 2016.
2. D. W. Bates, S. Saria, et al. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7):1123–1131, 2014.
3. W. Boag, D. Doss, et al. What's in a note? unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings*, 2018:26, 2018.
4. Y. Chechulin, A. Nazerian, et al. Predicting patients with high risk of becoming high-cost healthcare users in Ontario (Canada). *Healthcare Policy*, 9(3):68–79, 2014.
5. B. Chiu, G. Crichton, 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.
6. CIHI. Pan-Canadian Forum on High Users of Health Care, 2014.
7. E. Craig, C. Arias, and D. Gillman. Predicting readmission risk from doctors' notes. *arXiv:1711.10663*, 2017.
8. R. B. Deber and L. Kenneth. Handling the High Spenders : Implications of the Distribution of Health Expenditures for Financing Health Care. *American Political Science Association Annual Meeting*, pages 1–30, 2009.
9. D. Dligach and T. Miller. Learning Patient Representations from Text. *arXiv:1805.02096*, 2018.
10. S. Dubois, N. Romano, et al. Learning Effective Representations from Clinical Notes. *arXiv:1705.07025*, 2017.
11. L. Einav, A. Finkelstein, et al. Predictive modeling of U.S. health care spending in late life. *Science*, 360:1462–1465, 2018.
12. J. A. Fleishman and J. W. Cohen. Using information on clinical conditions to predict high-cost patients. *Health Services Research*, 45(2):532–552, 2010.
13. D. W. Frost, S. Vembu, et al. Using the Electronic Medical Record to Identify Patients at High Risk for Frequent Emergency Department Visits and High System Costs. *American Journal of Medicine*, 130(5):601.e17–601.e22, 2017.
14. G. Haixiang, L. Yijing, et al. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73(December):220–239, 2017.
15. D. A. Hanauer, Q. Mei, et al. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *Journal of Biomedical Informatics*, 55:290–300, 2015.
16. R. A. Hirth, T. B. Gibson, et al. New Evidence on the Persistence of Health Spending. *Medical Care Research and Review*, 72(3):277–297, 2015.
17. E. Homenauth, E. Graves, and L. Ishiguro. Examination of High-Cost Patients in Ontario. *International Journal of Population Data Science*, 3(3):359, 2018.
18. A. E. Johnson, T. J. Pollard, et al. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.
19. A. L. Kozyrskyj, L. Lix, and M. Dahl. *High-Cost Users of Pharmaceuticals : Who Are They ?* Manitoba Centre for Health Policy, 2005.
20. Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
21. C. Y. Li, D. Konomis, et al. Convolutional Neural Networks for Medical Diagnosis from Admission Notes. *arXiv:1712.02768*, 2017.
22. H. Liang, B. Y. Tsui, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Medicine*, 25(3):433–438, 2019.
23. J. Liu, Z. Zhang, and N. Razavian. Deep EHR: Chronic Disease Prediction Using Medical Notes. *arXiv:1808.04928*, 2018.

24. Y. Liu, T. Ge, et al. Exploiting Task-Oriented Resources to Learn Word Embeddings for Clinical Abbreviation Expansion. *arXiv:1804.04225*, 2018.
25. G. Louppe and M. Kumar. Scikit-optimize, 2018.
26. Y. Luo. Recurrent neural networks for classifying relations in clinical notes. *Journal of Biomedical Informatics*, 72:85–95, 2017.
27. T. Mikolov, K. Chen, et al. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781*, pages 1–12, 2013.
28. S. T. Moturu, W. G. Johnson, and H. Liu. Predicting future high-cost patients: A real-world risk modeling application. *Proc. BIBM 2007*, pages 202–208, 2007.
29. Ontario Hospital Association. Ideas and Opportunities for Bending the Health Care Cost Curve, 2010.
30. C. Paxton, A. Niculescu-Mizil, and S. Saria. Developing predictive models using electronic medical records: challenges and pitfalls. *Proc. AMIA Annual Symposium 2013*, 2013:1109–15, 2013.
31. B. Pedziński, P. Sowa, et al. Information and communication technologies in primary healthcare - Barriers and facilitators in the implementation process. *Studies in Logic, Grammar and Rhetoric*, 35(48):179–189, 2013.
32. J. Pennington, R. Socher, and C. Manning. Glove: Global Vectors for Word Representation. *Proc. EMNLP 2014*, pages 1532–1543, 2014.
33. A. M. Placona, R. King, and F. Wang. Longitudinal Clustering of High-cost Patients' Spend Trajectories : Delineating Individual Behaviors from Aggregate Trends. *AMIA Annu Symp Proc*, pages 907–915, 2018.
34. M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
35. S. Pyysalo, F. Ginter, et al. Distributional Semantics Resources for Biomedical Text Processing. *Proceedings of LBM*, pages 39–44, 2013.
36. S. Rais, A. Nazerian, et al. High-cost users of Ontario's healthcare services. *Healthcare Policy*, 9(1):44–51, 2013.
37. L. C. Rosella, T. Fitzpatrick, et al. High-cost health care users in Ontario, Canada: Demographic, socio-economic, and health status characteristics. *BMC Health Services Research*, 14(1), 2014.
38. B. A. Rosser, L. M. McCracken, et al. Concerns about medication and medication adherence in patients with chronic pain recruited from general practice. *Pain*, 152(5):1201–1205, 2011.
39. N. Shaw. The role of the professional association: A grounded theory study of Electronic Medical Records usage in Ontario, Canada. *IJIM*, 34(2):200–209, 2014.
40. X. Sun and W. Xu. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014.
41. S. Tamang, A. Milstein, et al. Predicting patient 'cost blooms' in Denmark: A longitudinal population-based study. *BMJ Open*, 7(1):1–10, 2017.
42. K. Tu, J. Klein-Geltink, et al. De-identification of primary care electronic medical records free-text data in Ontario, Canada. *BMC Medical Informatics and Decision Making*, 10(1), 2010.
43. K. Tu, T. F. Mitiku, et al. Evaluation of Electronic Medical Record Administrative data Linked Database (EMRALD). *American Journal of Managed Care*, 20(1):15–21, 2014.
44. Y. Wang, S. Liu, et al. A Comparison of Word Embeddings for the Biomedical Natural Language Processing. *arXiv:1802.00400*, 2018.
45. W. P. Wodchis, P. C. Austin, and D. Henry. A 3-year study of high-cost users of health care. *CMAJ*, 188(3):182–188, 2016.
46. W. P. Wodchis, K. Bushmeneva, et al. Guidelines on Person-Level Costing Using Administrative Databases in Ontario. *HSPRN*, 1, 2013.
47. C. J. Zook and F. D. Moore. High-Cost Users of Medical Care. *New England Journal of Medicine*, 302(18):996–1002, 1980.