JAMIA Open, 3(2), 2020, 160–166
doi: 10.1093/jamiaopen/ooaa022
Advance Access Publication Date: 15 July 2020
Brief Communications

Brief Communications

Generating contextual embeddings for emergency department chief complaints

David Chang [1], Woo Suk Hong [2] and Richard Andrew Taylor[2]

[1]Computational Biology and Bioinformatics Program, Yale University, New Haven, Connecticut, USA and [2]Department of Emergency Medicine, Yale School of Medicine, New Haven, Connecticut, USA

Corresponding Author: Richard Andrew Taylor, Department of Emergency Medicine, Yale School of Medicine, 464 Congress Ave, Ste 260 New Haven, CT 06519, USA (richard.taylor@yale.edu)

Received 29 January 2020; Revised 23 April 2020; Editorial Decision 11 May 2020; Accepted 14 May 2020

ABSTRACT

Objective: We learn contextual embeddings for emergency department (ED) chief complaints using Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art language model, to derive a compact and computationally useful representation for free-text chief complaints.
Materials and methods: Retrospective data on 2.1 million adult and pediatric ED visits was obtained from a large healthcare system covering the period of March 2013 to July 2019. A total of 355 497 (16.4%) visits from 65 737 (8.9%) patients were removed for absence of either a structured or unstructured chief complaint. To ensure adequate training set size, chief complaint labels that comprised less than 0.01%, or 1 in 10 000, of all visits were excluded. The cutoff threshold was incremented on a log scale to create seven datasets of decreasing sparsity. The classification task was to predict the provider-assigned label from the free-text chief complaint using BERT, with Long Short-Term Memory (LSTM) and Embeddings from Language Models (ELMo) as baselines. Performance was measured as the Top-k accuracy from $k = 1{:}5$ on a hold-out test set comprising 5% of the samples. The embedding for each free-text chief complaint was extracted as the final 768-dimensional layer of the BERT model and visualized using t-distributed stochastic neighbor embedding (t-SNE).
Results: The models achieved increasing performance with datasets of decreasing sparsity, with BERT outperforming both LSTM and ELMo. The BERT model yielded Top-1 accuracies of 0.65 and 0.69, Top-3 accuracies of 0.87 and 0.90, and Top-5 accuracies of 0.92 and 0.94 on datasets comprised of 434 and 188 labels, respectively. Visualization using t-SNE mapped the learned embeddings in a clinically meaningful way, with related concepts embedded close to each other and broader types of chief complaints clustered together.
Discussion: Despite the inherent noise in the chief complaint label space, the model was able to learn a rich representation of chief complaints and generate reasonable predictions of their labels. The learned embeddings accurately predict provider-assigned chief complaint labels and map semantically similar chief complaints to nearby points in vector space.
Conclusion: Such a model may be used to automatically map free-text chief complaints to structured fields and to assist the development of a standardized, data-driven ontology of chief complaints for healthcare institutions.

Key words: BERT, chief complaint, emergency medicine, machine learning, natural language processing

LAY SUMMARY

Patient care in the emergency department (ED) is guided by the patient's chief complaint, a concise statement regarding the patient's medical history, current symptoms, and reason for visit. Because chief complaints are often stored as free-text descriptions of varying length and quality, secondary use of chief complaint data in operational decisions and research has been impractical. Moreover, even when chief complaints are stored in a structured format in electronic health records, there exists no standard nomenclature on how they are categorized. To remedy this problem, we use Bidirectional Encoder Representations from Transformers, a state-of-the-art language model, on a dataset of 1.8 million free-text ED chief complaints to derive a numerical representation for chief complaints, called "contextual embeddings." We show that contextual embeddings accurately predict provider-assigned chief complaint labels and map chief complaints with similar meaning (eg "wheezing" and "breathing problem") to nearby points in vector space. The model with its associated embeddings may be used to automatically map free-text chief complaints to structured labels and to help derive a standardized dictionary of chief complaints for healthcare institutions.

BACKGROUND AND SIGNIFICANCE

Patient care in the emergency department (ED) is guided by the patient's chief complaint.[1–3] Collected during the first moments of the patient encounter, a chief complaint is a concise statement regarding the patient's medical history, current symptoms, and reason for visit. While a chief complaint can be represented in a structured format with predefined categories, it is often captured in unstructured, free-text descriptions of varying length and quality.[4] Moreover, even when chief complaints are stored in a structured format, there exists no standard nomenclature or guidance on how they should be categorized.[5,6] As a consequence, administrators and researchers frequently find chief complaint data difficult to use for downstream tasks such as quality improvement initiatives and predictive analytics.[7] Thus, the secondary use of chief complaint data in daily operational decisions and research has been hampered by its form and representation.

Advances in natural language processing (NLP) provide an opportunity to address many of the challenges of chief complaint data. Contextual language models such as Embeddings from Language Models (ELMo) and Bidirectional Encoder Representations from Transformers (BERT) are able to generate dense vector representations, or embeddings, of free-text data such that semantically similar words or documents are mapped to nearby points in vector space.[8–11] Such methods have been successfully applied in the medical domain.[12–19] Recent work has used contextual language models to generate embeddings for chief complaints in the primary care setting, using a small dataset of patient-generated text.[20]

Contextual embeddings for ED chief complaints have many desirable properties. They distill the complex information stored in free-text into a compact, numeric format while avoiding the data sparsity that results from converting categorical variables into dummy variables or from using traditional NLP models such as Term Frequency-Inverse Document Frequency (td-idf) and Bag of Words (BoW).[21–23] Moreover, a contextual embedding model trained specifically on ED triage notes stores appropriate information about chief complaints within the context of ED patient care, as opposed to word similarities within a large undifferentiated corpus.[12,17]

ED chief complaints have been an important part of many clinical decision support tools, including those for early sepsis detection, in-hospital mortality, patient disposition, and early ED return.[22–26] Contextual embeddings for ED chief complaints may facilitate incorporating free-text information into prediction models, as has been shown in models for in-patient readmission.[27,28] Contextual embeddings also enable us to calculate a numeric distance between any two chief complaints to determine their relatedness, or similarity, an elusive concept that has hampered outcomes research focusing on subgroups of chief complaints as well as quality improvement projects on short-term ED return.[29] Lastly, contextual embeddings may be used to derive a standardized, data-driven ontology of ED chief complaints that could be shared among healthcare institutions and research entities to minimize the variability in how chief complaint labels are assigned from ED to ED,[7,30,31] as has been suggested by recent work on the Hierarchical Presenting Problem Ontology (HaPPy).[5,21,32]

OBJECTIVES

In this study, we expand on prior work by applying BERT, a state-of-the-art language model, on a dataset of 1.8 million provider-generated free-text ED chief complaints from a healthcare system covering seven independent EDs.[9,19] We use Long Short-Term Memory (LSTM) and ELMo for baseline comparison. We show that the contextual embeddings generated by BERT accurately predict provider-assigned chief complaint labels and map semantically similar chief complaints to nearby points in vector space.

MATERIALS AND METHODS

Retrospective data on all adult and pediatric ED visits were obtained from a large healthcare system covering the period of March 2013 to July 2019, with a combined annual census of approximately 500 000 visits across seven independent EDs, three of which are community hospital-based. The centralized data warehouse for the electronic health record (EHR) system (Epic, Verona, WI) was queried for chief complaint data. This study was approved, and the informed consent process waived, by the Human Investigation Committee at the authors' institution (HIC 2000025236).

Chief complaint data in the Epic EHR are represented in two forms, a "presenting problem" that is a structured list of 1145 labels and a free-text chief complaint comment section in the form of an unstructured text box. The structured label system does not correspond to external nomenclatures such as SNOMED Clinical Terms. The free-text chief complaint is entered by the triage nurse at the moment of patient encounter, along with one or more presenting problems, which the nurse selects from a structured list after searching for a free-text term. We removed visits that did not contain both forms of chief complaint data, but examined the distribution of structured chief complaint labels without a comment section through categorical data analysis and Chi-Square distance metrics to determine that the reduced dataset was representative of the full dataset.[33] Visits that had been assigned more than one chief complaint label were treated as separate training instances.

Given the skewed distribution of chief complaint labels, where the 25 most common labels out of a total of 1145 account for roughly half of the dataset, chief complaint labels that comprised

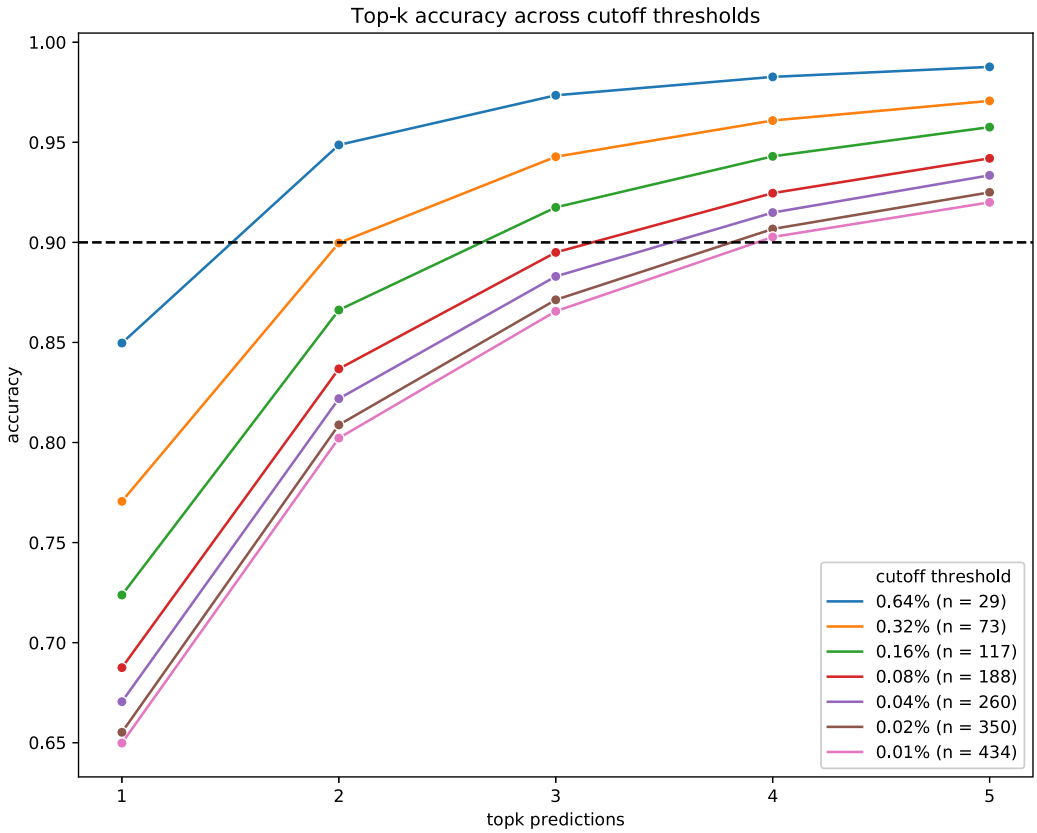

Figure 1. Model performance for Top-1 to Top-5 accuracy. Label-frequency cutoff thresholds are represented by colors. The accuracy increases drastically when taking into account the first few predictions. Dotted line shows 90% accuracy.

less than 0.01%, or 1 in 10 000, of all visits were excluded to ensure adequate training samples per label. The cutoff threshold was then incremented on a log scale to create seven datasets of decreasing sparsity (Supplementary Table S1). A full list of the chief complaint labels, along with their frequencies, are available in Supplementary Table S2.

Model training

For each of the seven datasets, all samples were randomly split into training (90%), validation (5%), and test (5%) sets. The classification task was to predict the provider-assigned label from the free-text chief complaint. Given the clinical nature of the dataset, we used a version of clinical BERT pretrained on the MIMIC corpus.[19] LSTM and ELMo were trained as baseline models on the largest dataset consisting of 434 labels.

Using the open source library PyTorch, we fine-tuned each clinical BERT model for three epochs on three GTX 1080 Ti GPUs. Each epoch on the full dataset took about an hour using *per_gpu_train_batch_size* of 144. Hyperparameter tuning beyond the default values for BERT fine-tuning did not yield noticeable gains in performance, with the test accuracies converging to the same range of values for any reasonable configuration. A *learning rate* of 1e−4 and *max_seq_length* of 64 were used. Sequences longer than max_seq_length were truncated. The implementation code is available at https://github.com/dchang56/chief_complaints. Notably, the repository also includes an easy-to-use script with instructions to generate predictions for custom chief complaint datasets.

For baseline comparison, we trained a bidirectional LSTM and ELMo using the AllenNLP framework. In both cases, the hidden dimension size was 512. The LSTM model was a one-layer bidirectional LSTM with GloVe embeddings, and the ELMo model was a two-layer biLSTM initialized with pretrained ELMo weights.

Performance

Performance was measured as the Top-k accuracy from $k = 1{:}5$ on a hold-out test set comprising 5% of the samples. Top-k accuracy is defined such that the model is considered to be correct if its top-k probability outputs contains the correct class label.

Error analysis

Having hundreds of potential labels with considerable semantic overlap (eg FACIAL LACERATION, LACERATION, HEAD LACERATION, FALL, FALL > 65) justifies taking into account the top few predictions rather than just the top 1. We hypothesized that the redundancy and noise in the label space would be responsible for the majority of the model's errors and *a priori* determined to examine through two-physician review a random sample of errors, as well as look at the most frequent kinds of mislabeling for common chief complaint labels.

Embedding visualization

The embedding for each free-text chief complaint was extracted as the final 768-dimensional layer of the BERT classifier. We took the mean of the embeddings across each chief complaint label and visualized the averaged, label-specific embeddings in a two-dimensional

space using t-SNE.[34] More specifically, the mean of the 768-dimensional embeddings across each chief complaint label was reduced to two dimensions using the *Rtsne* package (v. 0.15) in R with the following default hyperparameters: *initial_dims* = 50, *perplexity* = 30, *theta* = 0.5. To enhance readability of the figure, we limited the number of visualized labels to 188 by using a cutoff threshold of 0.08%. The *ggrepel* and *ggplot2* packages in R were used for plot generation. Clusters were determined via Gaussian mixture modeling with the optimal number selected by silhouette analysis.[35]

Table 1. Predictive performance by algorithm

	Algorithm	LSTM	ELMo	BERT
Full dataset (434 labels)	Top-1	0.63	0.63	0.65
	Top-2	0.77	0.78	0.80
	Top-3	0.84	0.85	0.87
	Top-4	0.88	0.88	0.90
	Top-5	0.90	0.90	0.92
Reduced dataset (188 labels)	Top-1	0.66	0.66	0.69
	Top-2	0.81	0.81	0.84
	Top-3	0.88	0.88	0.90
	Top-4	0.90	0.91	0.93
	Top-5	0.93	0.93	0.94

RESULTS

In the defined query time period, there were an initial 2 154 862 visits among 736 570 patients. 355 497 (16.4%) visits from 65 737 (8.9%) patients were removed for absence of either a structured or unstructured chief complaint. Among chief complaint labels, 43 of the 1145 labels were removed because of the absence of any visit with unstructured text. In comparison to the initial dataset, the chi-square distance metric for the histogram of the remaining chief complaint categories ($n = 1102$) was 0.005. For model training, an additional 668 labels comprising 25 143 (1.3%) visits were removed after filtering out labels that comprised less than 0.01%, or 1 in 10 000, of all visits, resulting in a total of 434 labels and 1 859 599 training instances.

The BERT models achieved increasing performance with higher label-frequency cutoff thresholds (Figure 1). BERT outperformed both LSTM and ELMo (Table 1). The BERT model yielded Top-1 accuracies of 0.65 and 0.69, Top-3 accuracies of 0.87 and 0.90, and

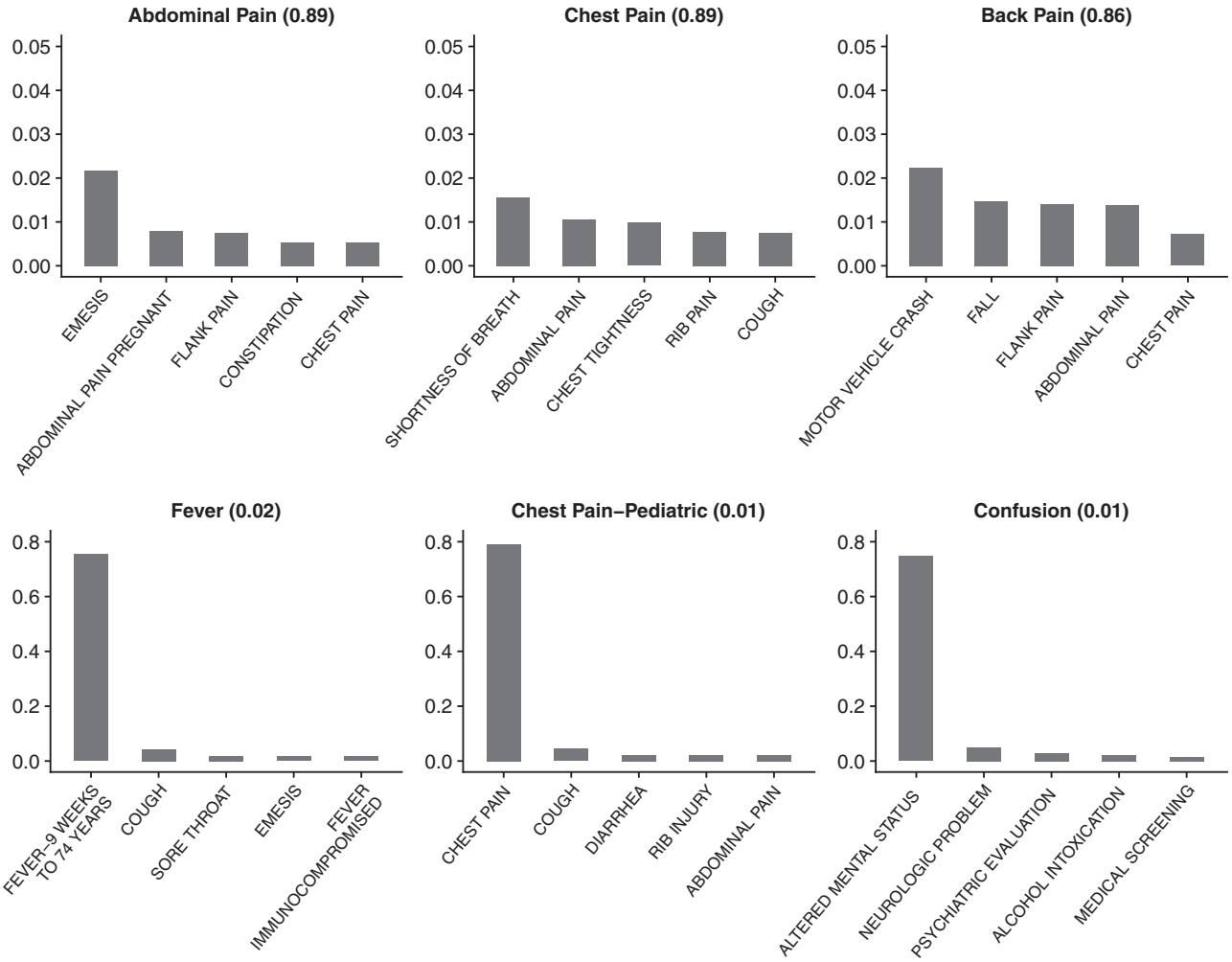

Figure 2. Common types of mislabeling for select chief complaint labels. Top row shows three of the most common chief complaint labels, with their accuracies shown within parentheses. Bottom row shows three chief complaint labels with lowest accuracies. X-axis shows the top five most common misclassifications in decreasing order. Y-axis shows frequency of error. Note that even for low performing chief complaint labels, a high percentage of errors are due to semantic overlap.

Table 2. Examples of chief complaints and their corresponding top-5 predictions

	Chief complaint	Top-5 predictions
Correctly classified at second prediction	"right third finger injured in door"	FINGER INJURY, *HAND PAIN*, HAND INJURY, FINGER PAIN, EXTREMITY LACERATION
	"pt comes to er with cc piece of plastic stuck to back of left ear from earing"	FOREIGN BODY IN EAR, *EAR PROBLEM*, EAR PAIN, OTALGIA, FOREIGN BODY
	"vomiting for days, increasing yesterday. pos home preg test on Saturday"	EMESIS, *EMESIS DURING PREGNANCY*, NAUSEA, ABDOMINAL PAIN PREGNANT, GI PROBLEM
	"both eyes swollen & itchy & tearing after his nap"	EYE SWELLING, *EYE PROBLEM*, EYE REDNESS, EYE PAIN, CONJUNCTIVITIS
	"fall at 0300 today, rt side weakness"	FALL, *FALL>65*, ALTERED MENTAL STATUS, NEUROLOGIC PROBLEM, WEAKNESS
Correctly classified at fifth prediction	"Felt like heart was pounding history of CABG. missed metoprolol for about 3 days."	PALPITATIONS, RAPID HEART RATE, TACHYCARDIA, IRREGULAR HEART BEAT, *CHEST PAIN*
	"2 weeks of sore throat, aches, dry cough. Denies intervention."	SORE THROAT, COLD LIKE SYMPTOMS, URI, COUGH, *FLU-LIKE SYMPTOMS*
	"fall down 5 stairs lace to right eyebrow"	FALL, FACIAL LACERATION, LACERATION, FALL>65, *HEAD LACERATION*
	"fever to 101, diarrhea, vomiting"	FEVER-9 WEEKS TO 74 YEARS, FEVER, EMESIS, ABDOMINAL PAIN, *FEVER-8 WEEKS OR LESS*
	"blister on back of foot."	BLISTER, FOOT PAIN, FOOT INJURY, FOOT SWELLING, *SKIN PROBLEM*

Top-5 accuracies of 0.92 and 0.94 on datasets comprised of 434 and 188 labels, respectively. Common types of mislabeling for the frequent chief complaint labels, as well as labels with the lowest accuracies, are shown in Figure 2. The interquartile range for Top-5 accuracies amongst the chief complaint labels was 74.0–92.3%.

Manual error analysis showed that many errors were due to the problem of redundancy and noise in the label space. In some cases, the predictions of the model were more suitable than the provider-assigned labels, as independently validated by physicians. We show 10 representative examples in Table 2 and provide a hundred random selection of errors in Supplementary Table S3.

The predictions are sorted in decreasing order of likelihood. The provider-assigned ground truth label is italicized. The examples highlight the problem of semantic overlap in the label space.

Figure 3 shows the t-SNE visualization of averaged embeddings for common chief complaint labels, clustered via Gaussian mixture modeling. Using the silhouette analysis, 15 was chosen to be the optimal number of clusters. A cutoff-threshold of 0.08% (ie 188 chief complaint labels) was used for readability in a two-dimensional space. t-SNE visualization for embeddings generated using LSTM and ELMo are shown in Supplementary Figures S4 and S5.

DISCUSSION

By applying BERT on a dataset of 1.8 million ED chief complaints from a healthcare system covering seven independent EDs, we derive contextual embeddings for chief complaints that accurately predict provider-assigned labels as well as map semantically similar chief complaints to nearby points in vector space.

Prior studies have derived embeddings for medical concepts, patient-to-provider messages, and primary care chief complaints.[12,18,20] We expand on prior work by using a large dataset of healthcare professional generated text, as opposed to patient-generated text, and by generating contextual embeddings for chief complaints within the emergency care setting. These embeddings may be instrumental in multiple downstream tasks, such as augmenting predictive performance of clinical decision support tools, calculating similarity measures between chief complaints to determine whether ED bounce-backs are due to a related cause,[29] or creating a standardized, data-driven ontology of chief complaints. Recently, much important work has been done to create a standardized ontology, namely, the Hierarchical Presenting Problem Ontology (HaPPy), which increased the likelihood of label assignment from a free-text chief complaint from 26.2% to 97.2% in one study.[5,21,32] Using such an ontology for training and testing purposes may present an opportunity for gold standard labels to be used to derive contextual embeddings.

Our study has several limitations. Our data come from a single healthcare system that uses an internal classification system for ED chief complaints, and our results may not be generalizable across EDs operating under different EHR systems. Moreover, certain conditions may be more likely to have structured chief complaint labels and by only training on that subset of patients, the model may have restricted applicability. Also, free-text chief complaints often list several comorbid signs and symptoms, making it difficult to choose a single ground truth label. This raises concerns about whether the prediction task should be set up as a multi-label classification task.

Another limitation is the noise inherent in the default set of chief complaint labels provided by our EHR. Of the 1145 default categories, 153 have one or no instance out of 1.87 million visits, while 472 account for 99% of the visits. Labels such as "Other" and "Medical" provide little to no information in an emergency care setting and restrict the applicability of the model. Some labels are synonyms (eg "Dyspnea" and "Shortness of breath"; "Otalgia" and "Ear pain"), while many more are hypo/hypernyms of one another (eg "Fall" and "Fall > 65"; "Migraine" and "Known dx migraine"). Such issues highlight the need to develop a principled and data-driven ontology for ED chief complaints. Despite the noise in the data, the model was able to learn a rich representation of chief complaints and generate reasonable predictions of their labels. In fact, many of the predictions that resulted in errors were more suitable

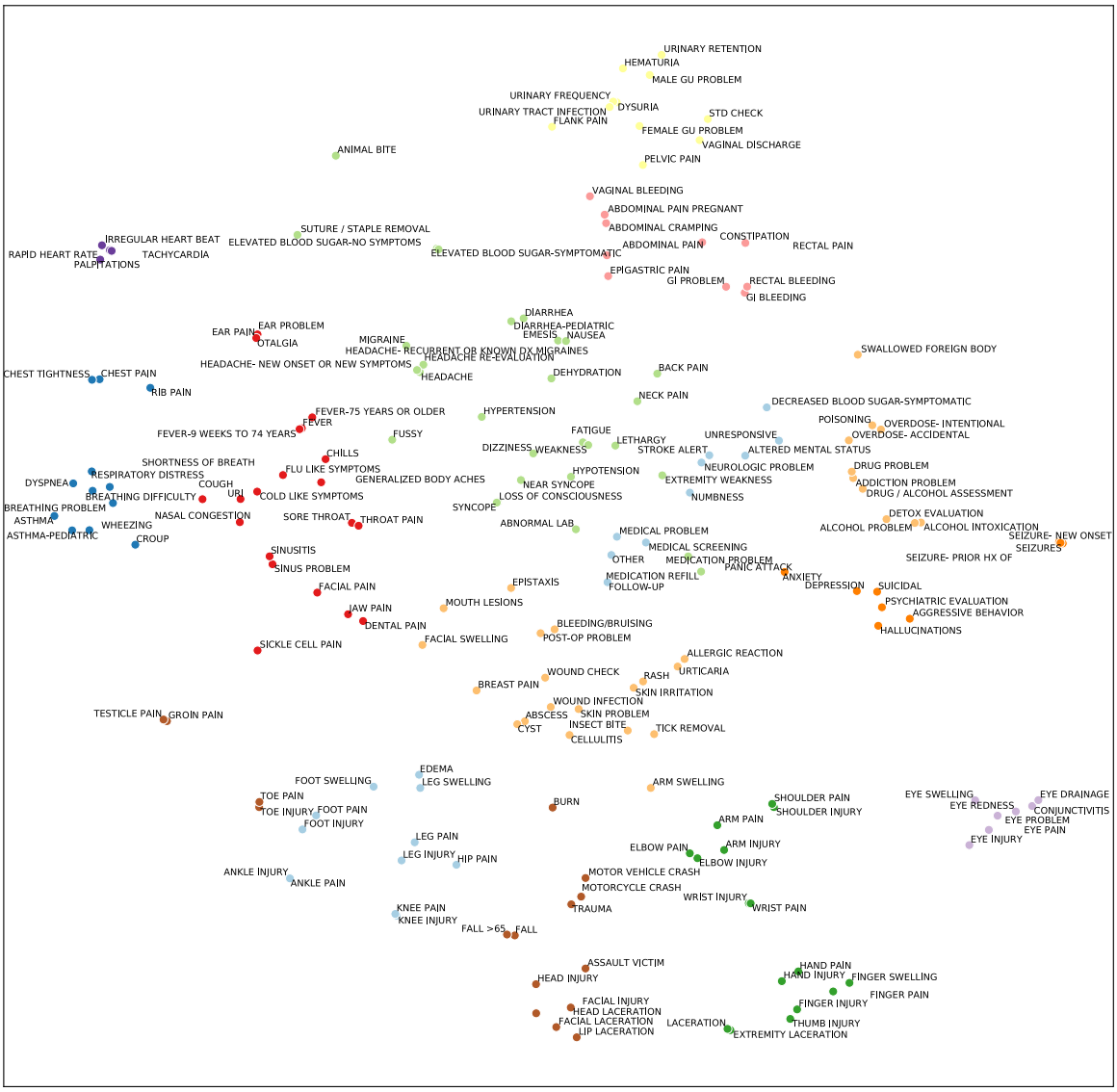

Figure 3. t-SNE visualization of averaged embeddings of common chief complaints. Embeddings for common chief complaints were grouped by their ground truth label, then their arithmetic mean visualized using t-SNE. The embeddings are distributed in a clinically meaningful way, with related concepts embedded close to each other and broader types of chief complaints clustered together. Note that t-SNE is a stochastic algorithm and, while it preserves local structure of the data, does not completely preserve its global structure. The text labels have been jittered to enhance readability. Colored groupings represent clusters as determined by gaussian mixture modeling.

than the ground truth labels, suggesting that the model did not overfit to the training data.

Finally, our model was trained only on free-text data, without any other patient information. Including non-textual patient data such as demographics, vital signs, and hospital usage statistics may improve performance, as shown in many prediction tasks.[25,26] Further studies are needed to assess the validity of these approaches.

CONCLUSION

The BERT language model was able to learn a rich representation of chief complaints and generate reasonable predictions of their labels despite the inherent noise in the label space. The learned embeddings accurately predicted provider-assigned chief complaint labels and mapped semantically similar chief complaints to nearby points in vector space. Such a model may be used to automatically map free-text chief complaints to structured fields and to derive a standardized, data-driven ontology of chief complaints for healthcare institutions.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

FUNDING

David Chang was supported by NIH Training Grant 5T15LM007056-33. Woo Suk Hong and Richard Andrew Taylor received no specific funding for this work.

AUTHORS' CONTRIBUTIONS

All authors contributed to the conception and design of the study. R.T. performed the acquisition and preprocessing of data. D.C., W.H., and R.T. performed the analysis of data, visualization, and interpretation of the results. D.C. and W.H. drafted the initial manuscript. D.C., W.H., and R.T. revised the final manuscript.

Conflict of interest statement. None declared.

REFERENCES

1. Griffey RT, Pines JM, Farley HL, *et al*. Chief complaint-based performance measures: a new focus for acute care quality measurement. *Ann Emerg Med* 2015; 65 (4): 387–95.
2. Mockel M, Searle J, Muller R, *et al*. Chief complaints in medical emergencies: do they relate to underlying disease and outcome? The Charité Emergency Medicine Study (CHARITEM). *Eur J Emerg Med* 2013; 20 (2): 103–8.
3. Mowafi H, Dworkis D, Bisanzo M, *et al*. Making recording and analysis of chief complaint a priority for global emergency care research in low-income countries. *Acad Emerg Med* 2013; 20 (12): 1241–5.
4. Haas S, Travers D, Tintinalli J, *et al*. Toward vocabulary control for chief complaint. *Acad Emerg Med* 2008; 15 (5): 476–82.
5. Horng S, Greenbaum NR, Nathanson LA, *et al*. Consensus development of a modern ontology of emergency department presenting problems—the Hierarchical Presenting Problem Ontology (HaPPy). *Appl Clin Inform* 2019; 10 (03): 409–20.
6. Aronsky D, Kendall D, Merkley K, *et al*. A comprehensive set of coded chief complaints for the emergency department. *Acad Emerg Med* 2001; 8 (10): 980–9.
7. Conway M, Dowling JN, Chapman WW. Using chief complaints for syndromic surveillance: a review of chief complaint based classifiers in North America. *J Biomed Inform* 2013; 46 (4): 734–43.
8. Peters ME, Neumann M, Iyyer M, *et al*. Deep Contextualized Word Representations. ArXiv180205365 Cs 22 March 2018; 14. http://arxiv.org/abs/1802.05365 (Accessed December 2019).
9. Devlin J, Chang M-W, Lee K, *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv181004805 Cs 24 May 2019; 13. http://arxiv.org/abs/1810.04805 (Accessed December 2019).
10. Le Q, Mikolov T. Distributed representations of sentences and documents In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. JMLR.org 2014. II–1188–II–1196. http://dl.acm.org/citation.cfm? id=3044805.3045025 (Accessed December 13, 2019).
11. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. USA: Curran Associates Inc. 2017: 6000–6010. http://dl.acm.org/citation.cfm?id=3295222.3295349 (Accessed December 13, 2019).
12. Choi Y, Chiu C-I, Sontag D. Learning low-dimensional representations of medical concepts. *AMIA Jt Summits Transl Sci Proc* 2016; 2016: 41–50.
13. Bai T, Chanda AK, Egleston BL, *et al*. Joint learning of representations of medical concepts and words from EHR data. *Proc IEEE Int Conf Bioinforma Biomed* 2017; 2017: 764–9.
14. Bai T, Chanda AK, Egleston BL, *et al*. EHR phenotyping via jointly embedding medical concepts and words into a unified vector space. *BMC Med Inform Decis Mak* 2018; 18 (S4): 123.
15. Beaulieu-Jones BK, Kohane IS, Beam AL. Learning contextual hierarchical structure of medical concepts with Poincairé embeddings to clarify phenotypes. *Pac Symp Biocomput Pac Biocomput* 2019; 24: 8–17.
16. Zhu H, Paschalidis IC, Tahmasebi A. Clinical Concept Extraction with Contextual Word Embedding. ArXiv181010566 Cs 26 November 2018; 14. http://arxiv.org/abs/1810.10566 (Accessed December 2019).
17. Si Y, Wang J, Xu H, *et al*. Enhancing clinical concept extraction with contextual embeddings. *J Am Med Inform Assoc* 2019; 26 (11): 1297–304.
18. Sulieman L, Gilmore D, French C, *et al*. Classifying patient portal messages using convolutional neural networks. *J Biomed Inform* 2017; 74: 59–70.
19. Alsentzer E, Murphy JR, Boag W, *et al*. Publicly Available Clinical BERT Embeddings. ArXiv190403323 Cs Published Online First: 20 June 2019. http://arxiv.org/abs/1904.03323 (Accessed December 13, 2019).
20. Valmianski I, Goodwin C, Finn IM, *et al*. Evaluating robustness of language models for chief complaint extraction from patient-generated text. ArXiv191106915 Cs 15 November 2019; 13. http://arxiv.org/abs/1911.06915 (Accessed December 2019).
21. Jernite Y, Halpern Y, Horng S, *et al*. Predicting chief complaints at triage time in the emergency department. In: NIPS 2013 Workshop on Machine Learning for Clinical Data Analysis and Healthcare 2013.
22. Horng S, Sontag DA, Halpern Y, *et al*. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One* 2017; 12 (4): e0174708.
23. Sterling NW, Patzer RE, Di M, *et al*. Prediction of emergency department patient disposition based on natural language processing of triage notes. *Int J Med Inf* 2019; 129: 184–8.
24. Taylor RA, Pare JR, Venkatesh AK, *et al*. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016; 23 (3): 269–78.
25. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLos One* 2018; 13 (7): e0201016.
26. Hong WS, Haimovich AD, Taylor RA. Predicting 72-hour and 9-day return to the emergency department using machine learning. *JAMIA Open* 2019; 2 (3): 346–52.
27. Xiao C, Ma T, Dieng AB, *et al*. Readmission prediction via deep contextual embedding of clinical concepts. *PLoS One* 2018; 13 (4): e0195024.
28. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. ArXiv190405342 Cs 11 April 2019; 15. http://arxiv.org/abs/1904.05342 (Accessed April 2020).
29. Rising KL, Victor TW, Hollander JE, *et al*. Patient returns to the emergency department: the time-to-return curve. *Acad Emerg Med* 2014; 21 (8): 864–71.
30. Chapman WW, Dowling JN, Wagner MM. Classification of emergency department chief complaints into 7 syndromes: a retrospective analysis of 527,228 patients. *Ann Emerg Med* 2005; 46 (5): 445–55.
31. Chapman W, Christensen L, Wagner M, *et al*. Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artif Intell Med* 2005; 33 (1): 31–40.
32. Greenbaum NR, Jernite Y, Halpern Y, *et al*. Improving documentation of presenting problems in the emergency department using a domain-specific ontology and machine learning-driven user interfaces. *Int J Med Inf* 2019; 132: 103981.
33. Yang L, Jin R. *Distance Metric Learning: A Comprehensive Survey*. Technical report, Department of Computer Science and Engineering, Michigan State University; 2006.
34. van der ML, Hinton G. Visualizing data using t-SNE. *Mach Learn* 2012; 87 (1): 33–605.
35. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987; 20: 53–65.