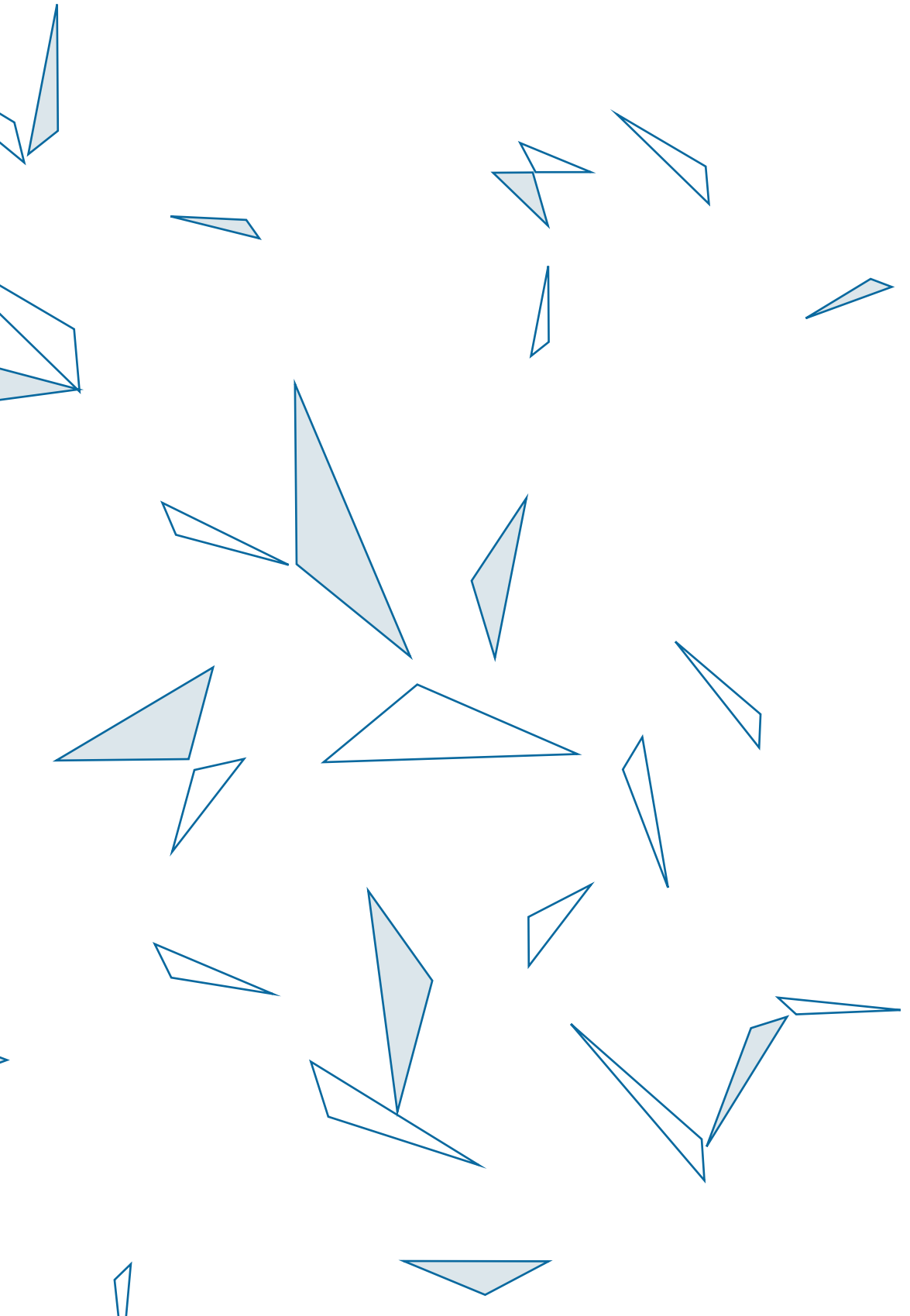
Applied machine learning in neurosurgical oncology

Senders, J.T.



8

Deep learning for natural language processing of free-text pathology reports

A comparison of learning curves

Joeky T. Senders, David J. Cote, Alireza Mehrtash, Robert Wiemann, William B. Gormley, Timothy R. Smith, Marike L.D. Broekman, Omar Arnaout

BMJ INNOVATIONS, 2020 VOL 6, ISSUE 4

Abstract

Introduction

Although clinically derived information could improve patient care, its full potential remains unrealized because most of it is stored in a format unsuitable for traditional methods of analysis, free-text clinical reports. Various studies have already demonstrated the utility of natural language processing algorithms for medical text analysis. Yet, evidence on their learning efficiency is still lacking. This study aimed to compare the learning curves of various algorithms and develop an open-source framework for text mining in healthcare.

Methods

Deep learning and regressions-based models were developed to determine the histopathological diagnosis of brain tumor patients based on free-text pathology reports. For each model, we characterized the learning curve and the minimal required training examples to reach the area under the curve (AUC) performance thresholds of 0.95 and 0.98.

Results

In total, we retrieved 7000 reports of 5242 brain tumor patients (2316 with glioma, 1412 with meningioma, and 1514 with cerebral metastasis). Conventional regression and deep learning-based models required 200-400 and 800-1500 training examples to reach the AUC performance thresholds of 0.95 and 0.98, respectively. The deep learning architecture developed in the current study required 100 and 200 examples, respectively, corresponding to a learning capacity that is two to eight times more efficient.

Conclusions

This open-source framework enables the development of high-performing and fast learning natural language processing models. The steep learning curve can be valuable for contexts with limited training examples (e.g., rare diseases and events or institutions with lower patient volumes). The resultant models could accelerate retrospective chart review, assemble clinical registries, and facilitate a rapid learning health care system.

Introduction

Clinically-derived patient information is generally increasing in volume and granularity with the expansion and improvement of online medical record systems.[1] Although analysis of this information allows for the generation of knowledge to improve future patient care, its full potential remains unrealized because most of it is stored in an unstructured, free-text format unsuitable for traditional methods of analysis. Manual review is necessary to extract and structure relevant patient information. As data sets continue to grow, however, manual chart review becomes increasingly inefficient, inconsistent, and prone to error.[2]

Various studies have already demonstrated the utility of automated methods for the processing and analysis of free-text clinical reports.[3–14] These algorithms have the potential to assist in structuring the immense stream of free-text clinical information produced on a day-to-day basis. However, they are generally evaluated on a static text corpus of clinical reports with a fixed number of training examples, thereby lacking evidence on their learning capacity (i.e., the required number of examples to train high-performing models).[3–14] Yet, learning efficiency is instrumental, as the availability of training examples can be limited, and their labelling can be time consuming or expensive.

In the current study, we aimed to compare the learning curves of various natural language processing approaches and develop an open-source framework for clinical text mining. Therefore, we have developed models that determine the histological diagnosis of brain tumor patients based on free-text pathology reports. Furthermore, we used various training samples to compare the efficiency of each algorithm's learning curve.

Methods

Participants

This study was conducted and reported according to the Transparent Reporting of a Multivariable Prediction Model of Individual Prognosis Or Diagnosis (TRIPOD) statement.[15] The Institutional Review Board of Brigham and Women's Hospital approved the current study and waived the need for informed consent due to its retrospective, observational design. We included all patients who underwent an operation at our institution for a histopathologically confirmed diagnosis of glioma, meningioma, or brain metastasis between January 2002 and July 2018. Patients were

identified through a departmental database that registers all neurosurgical patients who undergo an operation within our department. To retrieve the associated free-text pathology reports through which the diagnosis was made, we cross-linked the patient identification number and date of surgery with the pathology reports in our centralized institutional clinical data registry. These free-text pathology reports were used as input data for the natural language processing model. Manual annotations on the histopathological diagnosis were provided by a clinical reviewer with over 20 years of experience (R.W.). These annotations were used as labels for the target outcome in a binary fashion (diagnosis of interest versus other diagnoses). Some patients underwent multiple operations, thereby providing multiple pathology reports and associated diagnosis labels in the analysis. The total text corpus was split at the patient-level into a training, validation, and hold-out test set according to a 2:1:1 ratio. The test set was kept separate until the final performance evaluation. Differences in baseline characteristics between the training, validation, and hold-out test set were compared by means of the Chi-square test, analysis of variance (ANOVA), and Kruskal-Wallis test depending on the nature and distribution of the baseline characteristics.

Preprocessing

The algorithms used for this purpose can be classified into two broad categories, regression-based and neural network-based algorithms.[16,17] The regression-based algorithms utilized a bag-of-words/n-grams approach, thereby considering the relative frequency of words or adjacent word combinations in a document but ignoring their order.[17] Deep learning-based algorithms, on the other hand, modeled the order of the words and the semantic relationships among them, as well.[16,18]

The analysis of free-text pathology reports required both generic and approach-specific preprocessing steps. Pseudocode in a generalizable format is provided in Table 1. These preprocessing steps are required to compress the lexical content of free-text pathology reports to the most parsimonious representation and convert these reports to a numeric format that could be processed by a classification algorithm. Redundant or duplicate text (time, date, pathologist's signature, unnecessary white spaces etc.) was removed, and stemming was used to converge words with a similar lexical root.[19] For the regression-based algorithms, we used n-grams to assign unique value and meaning to adjacent word combinations.[20] The term frequency-inverse document frequency (TF-IDF) vectorization was used to convert each document into an array of numbers reflecting the relative frequency of these word or word combinations.[21] For the deep learning models, we tokenized and zero padded the documents to convert them to a numeric format with the same length.[22]

TABLE 1. Pseudocode of the current study in a generalizable format.

Step 1: Data importation and general preprocessing	A. Import data frame with three columns containing the patient identifier, group label, and original clinical report. B. In the original report column, subsequently a. remove all redundant information (date, time, physician's signature, white spaces between sections, punctuation between letters, and stop words) and transform all letters to lower case letters. b. remove all English stop words except 'no' and 'not' c. apply Porter stemming algorithm d. apply preprocessing steps C. Divide the stemmed reports at the patient-level into a training, validation, and test set in a 2:1:1 ratio.
Step 2: Hyperparameter optimization	A. Hyperparameter optimization by means of weighted bootstrapping on the training and validation set. Hyperparameter settings are further explained in Supplementary Table S1.
Step 3: Evaluate model performance on the hold-out test set.	A. Train final models with optimal hyperparameter settings on the training set with 100 bootstraps for each model and each training fraction. B. Compute the predicted probabilities in the residual hold-out test set. C. Calculate the pooled mean AUC with standard deviation for each model and training fraction. D. Plot the performance in AUC against the training sample size to visualize the resultant learning curve for each model.

Abbreviations: AUC=area under the receiver operating curve

Development of a natural language processing model

To compare the learning curve of the distinct approaches, model performance was evaluated for each diagnosis with varying samples ranging between 25 and 3000 reports of the training set. A bootstrapping procedure was utilized to optimize the hyperparameter settings. We used bootstrapping with replacement to draw random samples from the parent training set and trained 'naïve' algorithms with every iteration.[23] As such, all resultant models were solely trained on the randomly drawn sample while ignoring the rest of the parent training set. This bootstrapping procedure provided an estimate of performance for each hyperparameter setting by pooling the performance estimates of the distinct, sample-based models. To account for the higher variability in performance in the smaller samples and preserve the computational reproducibility of this study, the number of bootstraps was inversely weighted according to the sample size and comprised the integer division between the size of the total training set (n=3000) and the size of the training sample. For example, training with a sample of 25 reports was bootstrapped 120 times.

Logistic regression, least absolute shrinkage and selection operator (LASSO) regression, and deep learning models were developed and compared as classifiers.[24] For the regression-based models, the use of mono-, bi-, and trigrams, the size of the vocabulary

in the TF-IDF vectorizer, and l1 regularization were presented as hyperparameter settings. The following hyperparameters were optimized for the deep learning models: dimensionality of the embedding layer, dropout, kernel size, l1 regularization, l2 regularization, learning rate, max pooling window, number of convolutional layers, number of dense layers, number of filters in the convolutional layers, number of nodes in the dense layers, report length, type of optimizer, and vocabulary size of the tokenizer. Embedding and convolutional layers constitute the most instrumental layers in deep learning models used for natural language processing. In the embedding layer, a word can be represented by a vector of numbers instead of a single number.[16] These numbers represent the coordinates of a word in the embedding space and as such reflect the semantic relationships between individual words. Convolutional layers capture local interactions among nearby words by applying transformations with smaller one-dimensional filters on local regions of the input data.[25] Convolutional neural network (CNN) models are characterized by these layers and currently widely investigated because of their strong potential for image and text processing. Among the deep learning-based models, we therefore specifically compared the best performing CNN with non-convolutional neural network architectures. Explanations of the other hyperparameters are provided in Supplementary Table S1.

Evaluating final model performance

Training of the final models and evaluation on the residual hold-out test set was bootstrapped 100 times for each training fraction and model. The predicted outcomes of the natural language processing models constituted a probability of belonging to a histopathological class. Therefore, model performance was measured according to the area under the receiver operating characteristic curve (AUC).[26] The AUC is a measure of discrimination and represents the probability that an algorithm will rate a randomly selected case (i.e., category of interest) higher than a randomly selected non-case (i.e., all other cases). Model performance was pooled and weighted across the histopathological subclasses and plotted against the size of the training sample to construct each algorithm's learning curve. Based on these learning curves, we determined the minimal size of the training sample required to reach the AUC performance thresholds of >0.95 and >0.98. All models were trained and evaluated in Python version 3.6 (Python Software Foundation, http://www.python.org) using the Scikit-learn libraries. Figures for the incremental model performance were made in R version 3.3.3 (R Core Team, Vienna, Austria, https://cran.r-project.org/).[27]

Results

A total of 7000 pathology reports from 5242 patients were retrieved. Among these patients, 2316 (44.2%) were diagnosed with glioma, 1412 (26.9%) with meningioma, and 1514 with cerebral metastasis (28.9%). Baseline characteristics for the training, validation, and test set are provided in Table 2. A statistically significant difference (p=0.038) was found in the mean age across the cohorts. This difference was deemed to be of little clinical significance (57.0 years in the training set versus 56.4 years in the validation set) and likely the result of the large cohort sizes.

TABLE 2. Cohort characteristics

Patient Characteristics	Description	Training set (n = 2621)		Validation set (n = 1310)		Test set (n = 1311)		
		No.	%	No.	%	No.	%	*p*
Patient age	<50	728	27.8	423	32.3	394	30.1	0.060
	>70	519	19.8	236	18.0	251	19.1	
	50-70	1374	52.4	651	49.7	666	50.8	
	Mean ± SD	57.0 ± 14.4		56.4 ± 14.2		56.7 ± 14.7		0.038
Sex	Female	1474	56.2	720	55.0	738	56.3	0.716
	Male	1147	43.8	590	45.0	573	43.7	
Reports per patient	Median [IQR]	1 [1 - 1]		1 [1 – 1]		1 [1 – 1]		0.683
Histopathological diagnosis	Glioma	1142	43.6	574	43.8	600	45.8	0.697
	Meningioma	712	27.2	362	27.6	338	25.8	
	Metastasis	767	29.3	374	28.5	373	28.5	

Abbreviations: IQR=interquartile range; No.=sample size; p=p-value; SD=standard deviation

The neural network architecture developed in the current study, ClinicalTextMiner, demonstrated a steeper learning curve than regression-based models (Figure 1, Table 3) and other competing deep learning models (Figure 2). Regression-based algorithms required 200-400 and 800-1500 training examples to reach the AUC performance thresholds of 0.95 and 0.98, respectively. ClinicalTextMiner reached these thresholds with 100 and 200 examples, respectively, corresponding to a learning capacity that is two to eight times more efficient. The best performing CNN architecture reached the AUC performance threshold of 0.95 after training with at least 400 training examples and did not reach the performance threshold of 0.98. Furthermore, its

performance was less consistent compared to the other models as denoted by the larger standard deviations. Lastly, ClinicalTextMiner was the only model that reached the AUC performance threshold of >0.99, with 600 training examples.

TABLE 3. Incremental model performance according to the area under the receiver operating characteristic curve.

Sample size	Bootstrapped AUC			
	Deep learning-based models		Regression-based models	
	ClinicalTextMiner	CNN	Lasso regression	Logistic regression
25	0.794±0.052	0.751±0.085	0.698±0.106	0.507±0.029
50	0.901±0.038	0.835±0.062	0.850±0.070	0.583±0.109
75	0.943±0.023	0.885±0.052	0.904±0.033	0.715±0.096
100	0.958±0.016	0.904±0.038	0.929±0.022	0.785±0.078
150	0.977±0.008	0.925±0.028	0.949±0.014	0.890±0.034
200	0.983±0.005	0.936±0.024	0.959±0.010	0.918±0.019
250	0.986±0.003	0.943±0.024	0.965±0.009	0.929±0.014
300	0.987±0.002	0.945±0.020	0.968±0.008	0.944±0.011
400	0.988±0.002	0.954±0.015	0.973±0.006	0.950±0.009
500	0.989±0.002	0.956±0.015	0.977±0.004	0.961±0.009
600	0.990±0.001	0.956±0.014	0.981±0.003	0.969±0.006
800	0.991±0.001	0.962±0.012	0.983±0.003	0.974±0.005
1000	0.991±0.001	0.963±0.013	0.985±0.003	0.979±0.003
1200	0.991±0.001	0.963±0.013	0.986±0.002	0.981±0.002
1500	0.992±0.001	0.965±0.012	0.988±0.002	0.984±0.002
1800	0.992±0.001	0.965±0.013	0.989±0.002	0.985±0.002
2100	0.992±0.001	0.969±0.011	0.989±0.002	0.986±0.001
2500	0.992±0.001	0.964±0.013	0.990±0.001	0.988±0.001
3000	0.992±0.001	0.966±0.013	0.990±0.001	0.989±0.001

Abbreviations: AUC=area under the receiver operating characteristic curve; CNN=convolutional neural networks

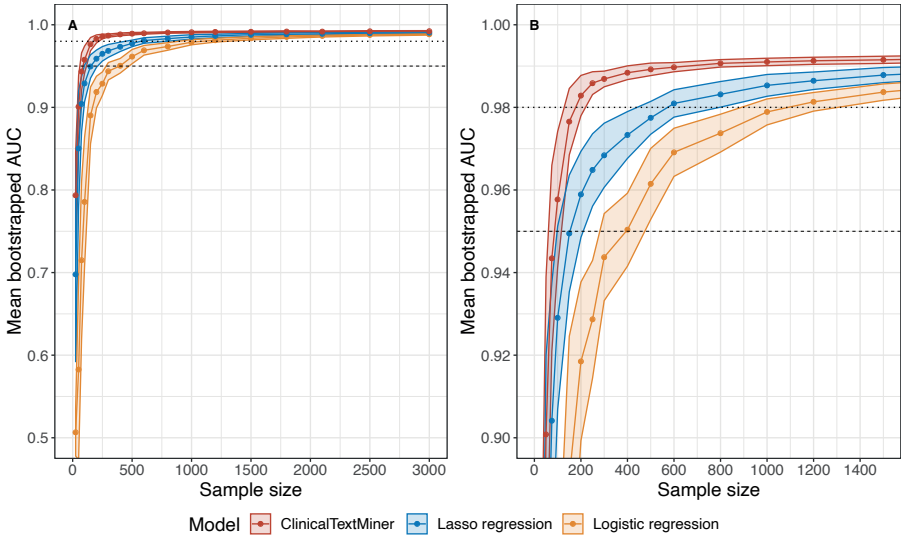


FIGURE 1. Incremental model performance comparing ClinicalTextMiner to regression-based algorithms according to the area under the receiver operating characteristic curve (A). Enlarged panel can be found on the right (B). The dashed line represents the 0.95 performance threshold and the dotted line represents the 0.98 performance threshold. Abbreviations: AUC= area under the receiver operating characteristic curve.

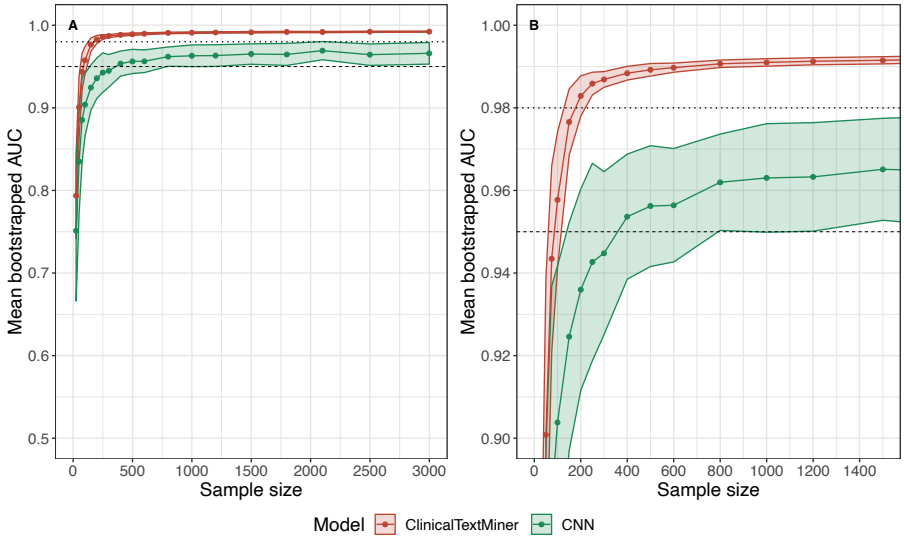


FIGURE 2. Incremental model performance comparing ClinicalTextMiner to the best performing convolutional neural network (CNN) model according to the area under the receiver operating characteristic curve (A). Enlarged panel can be found on the right (B). The dashed line represents the 0.95 performance threshold and the dotted line represents the 0.98 performance threshold. Abbreviations: AUC= area under the receiver operating characteristic curve; CNN=convolutional neural network.

Discussion

In this study, we compared the learning curves of various natural language processing techniques for clinical text mining. We identified a deep learning architecture that learns two to eight times more efficiently than competing deep learning and regressions-based approaches. The underlying source code has been released on a publicly accessible repository, thereby allowing for external application, validation, and optimization.

Few other groups have explored the use of deep learning for natural language processing of free-text clinical reports.[3–14] Although these studies demonstrate the strong potential of deep learning for clinical text mining, this has only been demonstrated on a static text corpus of clinical reports with a fixed number of training examples. The current body of literature therefore still lacks evidence on the learning capacity of natural language processing as it applies to clinical text mining. To the best of our knowledge, this is the first study that compares the learning curve of deep learning and other competing algorithms in their ability to process free-text clinical reports. Additionally, by removing the convolutional layer and combining the embedding layer with strong methods of regularization, we identified a model architecture that learns more efficiently compared to competing algorithms and even CNNs, thereby requiring less training examples to develop high performing clinical text mining models.

Limitations

Several limitations of the current study should be mentioned, which underline common barriers in natural language processing and machine learning modeling. The current models are trained on single institutional data and might not generalize well to data from external institutions, which may have different styles and routines in the language used in clinical reports. Instead of the resultant models, the underlying code pipeline was therefore made publicly accessible in order to promote the reproducibility and external generalizability of the current work. Labels were necessary for training and testing of the natural language processing models, and these were derived through manual chart review; however, manual chart review remains prone to error as well. In the current study, a trained clinical reviewer (R.W.) with over 20 years of experience has provided the required labels. The current study utilized a three-way classification problem (i.e., glioma, metastasis, or meningioma) to evaluate the learning curves. However, the number of diagnostic classes can vary widely dependent on the scope of interest, which could also impact the efficiency of the resultant learning curves.

Implications

Despite these limitations, we believe the current study provides valuable insight into the learning capacity of various methods for clinical text mining, as well as an open-source framework for developing these tools. In the current project, we used natural language processing to extract the histopathological diagnosis from free-text pathology reports of brain tumor patients. This application was chosen because the histopathological diagnosis constitutes the cornerstone for patient grouping in clinical research and day-to-day patient care. Currently, retrospective case identification is often based on ICD-9 codes and manual chart review. The accuracy of ICD-9 codes is questionable because they are developed for billing purposes and are often registered by non-medically trained assistants. Several studies have investigated the utility of ICD-9 codes, and all reported poor performance in terms of accuracy, sensitivity, specificity, or positive predictive value, depending on the diagnosis of interest.[28–31] Manual chart review, on the other hand, is labor intensive and drastically limits the speed, scale, and consistency of retrospective case identification.

It remains unclear why ClinicalTextMiner was able to outperform traditional (bag-of-words/n-gram) and novel methods (convolutional neural networks) of natural language processing. An explanation could be that the bag-of-words/n-gram approach focuses primarily on the relative frequency of certain word or adjacent word combinations in the text, whereas ClinicalTextMiner models the semantic properties and relations as well, due to the embedding layer at its base. The strong methods of regularization (i.e., pooling and dropout) could potentially avoid overfitting to statistical noise, which is introduced when analyzing semantic properties of words in addition to the relative word frequencies.

The model developed in the current study required only 100 (i.e., 30-35 per diagnostic group) and 200 training examples (i.e., 60-70 per diagnostic group) to reach the AUC performance thresholds of 0.95 and 0.98, respectively. This learning capacity can be instrumental as it allows for the development of high-performing clinical text mining models in applications with limited training examples. As such, models can also be developed in hospitals with lower patient volumes or utilized in the context of rare diseases and events. Furthermore, it reduces the workload on clinical experts who have to label each training example manually. In addition to the histopathological diagnosis, this open-source framework can guide the development of models to extract various clinical concepts from other report types, simply by providing the appropriate reports

and training labels. For example, labelling a radiology report with the size and location of the lesion allows for the development of a model that can extract radiological characteristics.[32–35]

Natural language processing could improve clinical research and patient care in several ways. The automatic nature could accelerate retrospective chart review to an unprecedented scale and allow for the assembly of large clinical registries. The deterministic nature can make data collection less subject to inter- and intra-reviewer inconsistencies but rather based on a consensus label from clinical experts. Lastly, by extracting patient characteristics and outcomes automatically, natural language processing could facilitate a health care system that continuously learns from clinically derived data. As such, these algorithms might assist in structuring the immense stream of free-text clinical information produced on a day-to-day basis. Models could, for example, be developed to automate trivial, administrative processes with low clinical impact (i.e., assigning the appropriate billing codes) or construct flagging systems and safety nets for matters of high-clinical impact (i.e., serious findings reported in diagnostic studies).

The current framework allows for the prediction of a single outcome and can be transformed to predict other outcomes as well. Developing distinct models for different outcomes in the same text corpus can, however, be computationally inefficient. After all, the underlying patterns and features learned from the text corpus might be generalizable to multiple outcomes and circumvent the need for a duplicate, time consuming training process. A solution could be to freeze the base model and repeat the training process only for the dense layers at the top of the network.[36] Another solution is emerging in the computational field and encompasses the development of deep learning algorithms for multiclass, multilabel classification problems. Future studies should therefore focus on evaluating the utility of these multiclass, multilabel algorithms for clinical text mining.[37] Furthermore, the learning curve of natural language processing algorithms remains relatively unexplored in the current literature and requires further investigation as well. The open-source framework developed in the current study could guide the construction of similar models for other clinical applications. Particularly, the efficient model architecture identified in the current study could be valuable for optimizing the hyperparameter settings in other deep learning-based natural language processing models. Lastly, future studies should not merely focus on the analytical challenges, but also on the implications of relying on automated methods for medical text analysis.

Conclusion

We developed an open-source natural language processing pipeline that can determine the histopathological diagnosis of brain tumor patients based on free-text pathology reports. A deep learning architecture was identified that learns two to eight times more efficiently than competing deep learning and regressions-based approaches. This framework could facilitate clinical research by providing an automatic and deterministic method for medical text analysis.

Supplementary material

Supplementary Table S1 and Figure S1 are available online at:

https://innovations.bmj.com/content/6/4/192

References

1. Evans RS. Electronic Health Records: Then, Now, and in the Future. *Yearb Med Inform*. 2016;(Suppl 1):S48-S61. doi:10.15265/IYS-2016-s006
2. Matt V, Matthew H. The retrospective chart review: important methodological considerations. *J Educ Eval Health Prof*. 2013;10. doi:10.3352/jeehp.2013.10.12
3. Bao Y, Deng Z, Wang Y, et al. Using Machine Learning and Natural Language Processing to Review and Classify the Medical Literature on Cancer Susceptibility Genes. *JCO Clinical Cancer Informatics*. 2019;(3):1-9. doi:10.1200/CCI.19.00042
4. Senders JT, Karhade AV, Cote DJ, et al. Natural Language Processing for Automated Quantification of Brain Metastases Reported in Free-Text Radiology Reports. *JCO Clinical Cancer Informatics*. 2019;In Press.
5. Shi X, Yi Y, Xiong Y, et al. Extracting entities with attributes in clinical text via joint deep learning. *J Am Med Inform Assoc*. doi:10.1093/jamia/ocz158
6. Spandorfer A, Branch C, Sharma P, et al. Deep learning to convert unstructured CT pulmonary angiography reports into structured reports. *Eur Radiol Exp*. 2019;3(1):37. doi:10.1186/s41747-019-0118-1
7. Chen P-H, Zafar H, Galperin-Aizenberg M, Cook T. Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports. *J Digit Imaging*. 2018;31(2):178-184. doi:10.1007/s10278-017-0027-x
8. Bacchi Stephen, Oakden-Rayner Luke, Zerner Toby, Kleinig Timothy, Patel Sandy, Jannes Jim. Deep Learning Natural Language Processing Successfully Predicts the Cerebrovascular Cause of Transient Ischemic Attack-Like Presentations. *Stroke*. 2019;50(3):758-760. doi:10.1161/STROKEAHA.118.024124
9. Leyh-Bannurah S-R, Tian Z, Karakiewicz PI, et al. Deep Learning for Natural Language Processing in Urology: State-of-the-Art Automated Extraction of Detailed Pathologic Prostate Cancer Data From Narratively Written Electronic Health Records. *JCO Clinical Cancer Informatics*. 2018;(2):1-9. doi:10.1200/CCI.18.00080
10. Taggart M, Chapman WW, Steinberg BA, et al. Comparison of 2 Natural Language Processing Methods for Identification of Bleeding Among Critically Ill Patients. *JAMA Netw Open*. 2018;1(6). doi:10.1001/jamanetworkopen.2018.3451
11. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology*. 2019;291(1):196-202. doi:10.1148/radiol.2018180921
12. Kehl KL, Elmarakeby H, Nishino M, et al. Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. *JAMA Oncol*. Published online July 25, 2019. doi:10.1001/jamaoncol.2019.1800
13. Wei Q, Ji Z, Li Z, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *J Am Med Inform Assoc*. Published online May 28, 2019. doi:10.1093/jamia/ocz063
14. He T, Puppala M, Ezeana CF, et al. A Deep Learning-Based Decision Support Tool for Precision Risk Assessment of Breast Cancer. *JCO Clin Cancer Inform*. 2019;3:1-12. doi:10.1200/CCI.18.00121
15. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594. doi:10.1136/bmj.g7594
16. Wu S, Roberts K, Datta S, et al. Deep learning in clinical natural language processing: a methodical review. *J Am Med Inform Assoc*. 2020;27(3):457-470. doi:10.1093/jamia/ocz200
17. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;8. doi:10.1186/s13643-019-1074-9

18. Gonçalves S, Cortez P, Moro S. A deep learning classifier for sentence classification in biomedical and computer science abstracts. *Neural Computing and Applications*. Published online 2019. doi:10.1007/s00521-019-04334-2
19. Cai T, Giannopoulos AA, Yu S, et al. Natural Language Processing Technologies in Radiology Research and Clinical Applications. *Radiographics*. 2016;36(1):176-191. doi:10.1148/rg.2016150080
20. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. 2011;18(5):544-551. doi:10.1136/amiajnl-2011-000464
21. Zhang W, Yoshida T, Tang X. TFIDF, LSI and multi-word in information retrieval and text categorization. In: *2008 IEEE International Conference on Systems, Man and Cybernetics*. ; 2008:108-113. doi:10.1109/ICSMC.2008.4811259
22. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging*. 2018;9(4):611-629. doi:10.1007/s13244-018-0639-9
23. Henderson AR. The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data. *Clin Chim Acta*. 2005;359(1-2):1-26. doi:10.1016/j.cccn.2005.04.002
24. Ranstam J, Cook JA. LASSO regression. *BJS*. 2018;105(10):1348-1348. doi:10.1002/bjs.10895
25. Zola P, Cortez P, Ragno C, Brentari E. Social Media Cross-Source and Cross-Domain Sentiment Classification. *International Journal of Information Technology & Decision Making (IJITDM)*. 2019;18(05):1469-1499.
26. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2
27. Modern Optimization with R | Paulo Cortez | Springer. Accessed April 6, 2020. https://www.springer.com/gp/book/9783319082622
28. Labovitz DL. Accuracy and yield of ICD-9 codes for identifying children with ischemic stroke. Published online November 22, 2018. Accessed November 22, 2018. http://n.neurology.org/content/accuracy-and-yield-icd-9-codes-identifying-children-ischemic-stroke
29. Pimentel MA, Browne EN, Janardhana PM, et al. Assessment of the Accuracy of Using ICD-9 Codes to Identify Uveitis, Herpes Zoster Ophthalmicus, Scleritis, and Episcleritis. *JAMA Ophthalmol*. 2016;134(9):1001-1006. doi:10.1001/jamaophthalmol.2016.2166
30. Guevara RE, Butler JC, Marston BJ, Plouffe JF, File TM, Breiman RF. Accuracy of ICD-9-CM Codes in Detecting Community-acquired Pneumococcal Pneumonia for Incidence and Vaccine Efficacy Studies. *Am J Epidemiol*. 1999;149(3):282-289. doi:10.1093/oxfordjournals.aje.a009804
31. Goldstein Larry B. Accuracy of ICD-9-CM Coding for the Identification of Patients With Acute Ischemic Stroke. *Stroke*. 1998;29(8):1602-1604. doi:10.1161/01.STR.29.8.1602
32. Tang R, Ouyang L, Li C, et al. Machine learning to parse breast pathology reports in Chinese. *Breast Cancer Res Treat*. 2018;169(2):243-250. doi:10.1007/s10549-018-4668-3
33. Imler TD, Morea J, Kahi C, et al. Multi-Center Colonoscopy Quality Measurement Utilizing Natural Language Processing. *The American Journal of Gastroenterology*. 2015;110(4):543-552. doi:10.1038/ajg.2015.51
34. Imler TD, Morea J, Kahi C, Imperiale TF. Natural Language Processing Accurately Categorizes Findings From Colonoscopy and Pathology Reports. *Clinical Gastroenterology and Hepatology*. 2013;11(6):689-694. doi:10.1016/j.cgh.2012.11.035
35. Jouhet V, Defossez G, Burgun A, et al. Automated Classification of Free-text Pathology Reports for Registration of Incident Cases of Cancer. *Methods Inf Med*. 2012;51(03):242-251. doi:10.3414/ME11-01-0005
36. Shin H-C, Roth HR, Gao M, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*. 2016;35(5):1285-1298. doi:10.1109/TMI.2016.2528162

37. Gargiulo F, Silvestri S, Ciampi M. Deep Convolution Neural Network for Extreme Multi-label Text Classification: In: *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications; 2018:641-650. doi:10.5220/0006730506410650