Radiology

Deep Learning to Classify Radiology Free-Text Reports[1]

Matthew C. Chen, MS
Robyn L. Ball, PhD
Lingyao Yang, PhD
Nathaniel Moradzadeh, MD
Brian E. Chapman, PhD
David B. Larson, MD, MBA
Curtis P. Langlotz, MD, PhD
Timothy J. Amrhein, MD
Matthew P. Lungren, MD, MPH

Purpose: To evaluate the performance of a deep learning convolutional neural network (CNN) model compared with a traditional natural language processing (NLP) model in extracting pulmonary embolism (PE) findings from thoracic computed tomography (CT) reports from two institutions.

Materials and Methods: Contrast material–enhanced CT examinations of the chest performed between January 1, 1998, and January 1, 2016, were selected. Annotations by two human radiologists were made for three categories: the presence, chronicity, and location of PE. Classification of performance of a CNN model with an unsupervised learning algorithm for obtaining vector representations of words was compared with the open-source application PeFinder. Sensitivity, specificity, accuracy, and F1 scores for both the CNN model and PeFinder in the internal and external validation sets were determined.

Results: The CNN model demonstrated an accuracy of 99% and an area under the curve value of 0.97. For internal validation report data, the CNN model had a statistically significant larger F1 score (0.938) than did PeFinder (0.867) when classifying findings as either PE positive or PE negative, but no significant difference in sensitivity, specificity, or accuracy was found. For external validation report data, no statistical difference between the performance of the CNN model and PeFinder was found.

Conclusion: A deep learning CNN model can classify radiology free-text reports with accuracy equivalent to or beyond that of an existing traditional NLP model.

Online supplemental material is available for this article.

[1] From the Department of Radiology, Stanford University School of Medicine, Stanford University Medical Center, 725 Welch Rd, Room 1675, Stanford, Calif 94305-5913 (M.C.C., N.M., D.B.L., C.P.L., M.P.L.); Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, Calif (R.L.B., L.Y.); Department of Bioinformatics, University of Utah Medical Center, Salt Lake City, Utah (B.E.C.); and Department of Radiology, Duke University Medical Center, Durham, NC (T.J.A.). Received May 18, 2017; revision requested July 24; revision received July 25; accepted August 11; final version accepted September 14. **Address correspondence to** M.P.L. (e-mail: *mlungren@stanford.edu*).

Supported by Philips (1199247-130-UBGKW) and Stanford Child Health Research Institute (UL1 TR001085).

Medical imaging reports contain valuable diagnostic information. The majority of reports that accompany the imaging examination are composed of free-text narration. This format is the principal reason that most information in medical imaging reports is unstructured and remains inaccessible to automated analysis. Structured information from the text of imaging reports can be applied in many settings, including input for clinical decision support, monitoring for appropriate use of medical imaging, unbiased identification of disease cohorts, and labeling of medical images for computer vision applications (1–7).

Natural language processing (NLP) has shown promise in automating the classification of free narrative text. Examples of this include Medical Language Extraction and Encoding System (MedLEE), which relies on controlled vocabulary and grammatical rules to convert free text into a structured database (2,4). Another similar program called Lexicon Mediated Entropy Reduction (LEXIMER) was used to process radiology reports to identify critical recommendations based on a similar methodology (1). An NLP system for pulmonary embolism (PE) classification called PeFinder (*https://code.google.com/p/negex*) uses defined lexical cues and context terms with high accuracy (8). However, one of the main drawbacks of these and other similar existing NLP techniques is the relatively high burden of development, including domain-specific feature engineering, complex annotations, and laborious coding for specific tasks.

Rapid advances in computing power have prompted the rise of a new generation of machine learning methods, broadly referred to as deep learning. A common deep learning technique known as the convolutional neural network (CNN) has been effective in visual-object and speech-recognition tasks, achieving performance beyond other more labor-intensive methods (5). Further, CNNs have shown promise in the classification of medical imaging and diagnostic tasks (6,7,9,10). Not limited to images alone, CNNs may also be an extremely powerful tool for free-text annotation; the principal advantage is that CNNs can achieve optimal performance without the drawbacks of traditional NLP approaches because they do not require the development of dictionary-based libraries, grammatical feature definitions, concept codes, sentence-level annotation, or the generation of predefined terms. In other words, CNNs have been shown to identify the important words and phrases in the report text organically, without prior input as to which words or phrases were important for the labeling task (8,11,12). It is possible that, if successfully applied to medical imaging text, CNNs could quickly and accurately classify text with a low development cost and high generalizability.

The purpose of our study was to evaluate the performance of a deep learning CNN model compared with traditional NLP models in extracting PE findings from thoracic computed tomography (CT) reports from two institutions. By successfully automating the classification of imaging reports, this work may enable large-scale CNN annotation of free text in medical imaging reports.

Advances in Knowledge

- A deep learning convolutional neural network (CNN) model for natural language processing (NLP) can classify radiology free-text reports with accuracy equivalent to or beyond that of an existing traditional NLP model and attained an accuracy of 99% and an area under the curve value of 0.97 for determining the presence of pulmonary embolism in contrast material–enhanced chest CT reports.
- The CNN model had a statistically significant larger F1 score than did PeFinder (the current state-of-the-art machine learning classifier), representing a new benchmark in the classification of radiology free-text reports.

Implication for Patient Care

- This highly accurate, generalizable, deep learning–based automated software application for automated classification of radiology free-text reports could be made available for a variety of applications, including diagnostic surveillance, cohort building, quality assessment, labels for computer vision data, and clinical decision support services.

Materials and Methods

Financial Support

Financial support for this project was provided by grants from Philips (1199247-130-UBGKW) and the Stanford Child Health Research Institute (Stanford NIH-NCATS-CTSA grant no. UL1 TR001085). The authors had control of all data and the information submitted for publication, and none are employees of or consultants for third parties related to this research.

Clinical Data

We obtained radiology reports for contrast material–enhanced CT examinations of the chest performed between July 1995 and April 2016. We identified

https://doi.org/10.1148/radiol.2017171115

Content code: IN

Radiology 2018; 000:1–8

Abbreviations:
CI = confidence interval
CNN = convolutional neural network
NLP = natural language processing
PE = pulmonary embolism

Author contributions:
Guarantors of integrity of entire study, M.C.C., M.P.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, M.C.C., C.P.L., T.J.A., M.P.L.; clinical studies, T.J.A.; experimental studies, M.C.C., N.M., B.E.C.; statistical analysis, M.C.C., R.L.B., L.Y.; and manuscript editing, all authors

Conflicts of interest are listed at the end of this article.

Figure 1

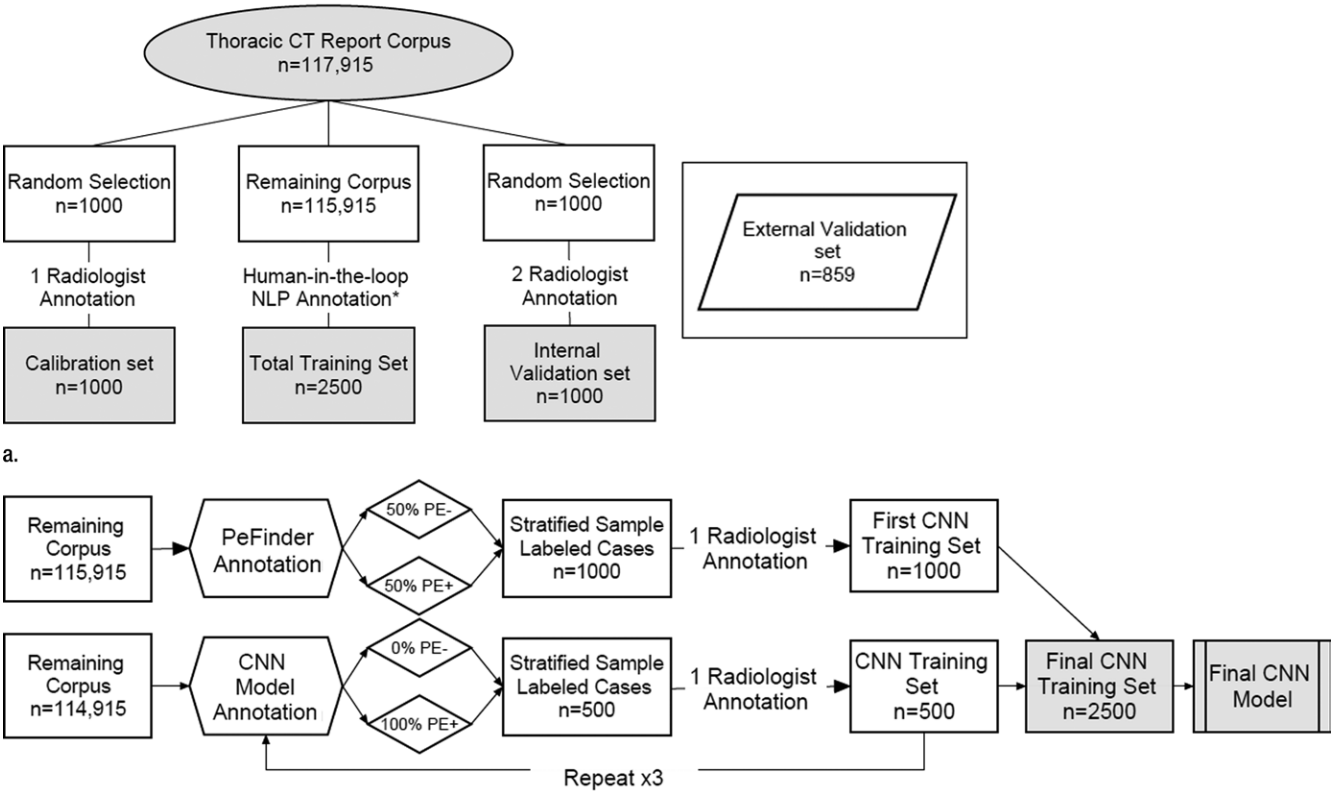

Figure 1: **(a)** Schematic for deriving training, calibration, and internal validation reports. No training reports were used in calibration or internal validation sets. External validation set was annotated and provided by developers of PeFinder; these studies were pulled from corpus of thoracic CT reports at a different institution. Reports from external validation set were not included in training, calibration, or internal validation set. **(b)** Schematic for "human-in-the-loop" CNN model training. Initial CNN model was trained on a stratified sample of remaining corpus labeled by PeFinder to select training data. These 1000 annotated reports were then annotated by a radiologist and used to train the first CNN model because PE-positive cases are rare (approximately 10%) in overall corpus and are important to include in a high proportion for radiologist annotation for training the model. Afterward, each trained CNN model was used iteratively to select 500 PE-positive cases at a time for radiologist annotation and CNN model training. Final CNN model is based on total of 2512 document-level radiologist-annotated reports.

chest CT reports by using radiology procedure codes. The impressions were extracted from the original reports by searching for the "impression" keyword that signified the beginning of the impression section and by using simple heuristic techniques for determining the end of the impression. All examinations were deidentified in a fully Health Insurance Portability and Accountability–compliant manner and acquisition, and processing of data were approved by the institutional review board of Stanford University.

Report Annotation

The CNN model used was based on TensorFlow, which is a freely available open-source deep learning library (*https://www.arxiv.org/abs/1605.08695*). A total of 2500 annotated reports were used to train the CNN model (Fig 1). The first training set comprised 1000 reports; these were a stratified random sample (roughly 500 reports were PE positive and 500 were PE negative) selected from the study population after first labeling the entire corpus by using PeFinder. Because PE is rare, we used the labels of PeFinder as a starting point, but each report was annotated by a radiologist and the radiologist's annotation label (not the PeFinder label) was considered the "truth" in the training set. With this starting set of 1000 reports, the CNN model was used to classify all reports. Next, three iterations of the described training method for labeling and annotating were used for the final CNN model; no significant improvement was observed with further iterations. Each iteration consisted of selecting a random sample of 500 reports that the CNN model labeled as PE positive. A radiologist (M.L., with 8 years of experience) then annotated these reports, and the radiologist's labels were treated as the truth in the next round of training (Fig 1b).

To create the calibration and internal validation set, 2000 reports were randomly selected from the study population and split evenly (Fig 1a). The calibration set was used to tune performance in the model-building process,

whereas the internal validation set was used to evaluate the performance in the final model. A single radiologist (M.L.) annotated all of the reports in the calibration set and two radiologists (M.L. and N.M., with 3 years of experience) independently annotated each report in the internal validation set; any discrepancy was resolved in another round by the senior radiologist (M.L.). The interrater reliability κ scores were also calculated by the two radiologists. No training reports were used in either the calibration or internal validation sets.

The external validation set comprised contrast-enhanced thoracic CT reports from the University of Pittsburg Medical Center, providing a means to evaluate the generalizability of the CNN model to accurately label reports for which it was not trained. These data were originally used to design and test PeFinder (8). We replicated the published results of the performance of the PeFinder model and then compared it with the performance of the CNN model.

All document-level annotations by human radiologists were made for three categories: presence of PE, present or absent; chronicity of PE, acute or chronic; and location of PE, central or subsegmental only. If a PE was present in the report, then it was considered a positive study for PE. Chronicity was labeled as either acute or chronic based on the description in the text report. In the setting of acute-on-chronic PE, or "mixed" chronicity, the report was to be labeled as acute. The "subsegmental only" label was used in cases in which the PE was described as subsegmental and did not include more central locations. Interrater percentage agreement and κ scores were calculated.

CNN Model Framework

The specific parameters for the CNN model were obtained from a previously published model called GloVe (Global Vectors for Word Representation). GloVe is an unsupervised learning algorithm for obtaining vector representations for words by using a log-bilinear model with a weighted least-squares objective based on the ratios of word-word co-occurrence probabilities (13). The main intuition underlying the model is the simple observation that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning. Briefly, the model automatically converts each word, including punctuation and other special characters, in a given report to a dense vector representation (14). GloVe model vectors are trained on the nonzero entries of a global word-word co-occurrence matrix, which tabulates how frequently words co-occur with one another in a given corpus. Words that did not match the GloVe corpus were converted to an unknown word, which was assigned to a random vector.

The truncated report impression section from each radiology report was used as the input, and the resulting word embeddings were padded with zero vectors at the end to ensure that all training examples had the same size. These embeddings were then processed through a convolutional layer with 200 filters and a window size of 10 words; hyperparameters were chosen by optimizing the objective function in the calibration set. The concept of word embeddings can be described as a mathematical representation of a word that can be used to derive similarities between words, as well as other relations. The typical example for word embeddings is that they allow one to perform math such as this: king − queen = man − woman, because the difference between the word embeddings is similar, and therefore king − queen + woman = man (15).

The CNN model architecture in our study was identical to Kim (14), and more detail on choice of parameters and model schematic can be found in the Appendix E1 (online). In brief, a report impression with 250 words would be converted into an element matrix of 300 × 250, where each column represents the embedding for a corresponding word. The matrix would then be padded with zero vectors to create the final input size of 300 × 300. A single filter would be represented as a 300 × 10 matrix of real-valued numbers. The convolutional layer performs element-wise matrix multiplications in a sliding-window manner to generate output sent to the next layer. One natural interpretation of this process is scanning a report for relevant local structure or phrases, which is learned in the training process. This layer was followed by a maximum pooling layer, which takes the maximum across all values of a given filter. Dropout, a regularization method to prevent overfitting of the data by randomly zeroing out connections during training time, was then used on the output. The result was fed into a fully connected output layer, followed by a sigmoid layer to convert the raw scores to probabilities. The CNN model was trained on report-labeling tasks for PE (present or absent, acute or chronic, and central or subsegmental only) on a single model with three output nodes.

Visualization

One of the drawbacks of a CNN model is that the rationale for the labeling decisions is not apparent, commonly referred to as a "black box." However, methods have been developed to visualize the important features of a given input for the labeling output decisions to better understand the information used by a deep learning model in a natural language classification task (16). To elucidate the words that led to the CNN result, we used the L1 norm of the partial derivative of the loss function with respect to each input variable (word vector) as the importance score of each word.

Statistical Analysis and Comparison to PeFinder

We compared the performance of the CNN model to the previously published and freely available PeFinder model (8). We tested both PeFinder and the CNN model in the internal and external validation sets, as described previously. The external validation set was used to determine how the CNN model would perform on a different organization's data where no labeled training set was available to the model. To determine if the performance measures were statistically different between the CNN model and PeFinder, we calculated 95% bootstrap confidence intervals (CIs) of the difference, as follows: We calculated the performance measures for the CNN model and PeFinder on 10000 bootstrap

Comparison of PeFinder and the CNN Model for "PE Present" and "Acute PE" Classification Tasks

	PE Present		Acute PE	
Parameter	Internal Validation	External Validation	Internal Validation	External Validation
F1 score				
CNN model	0.938	0.891	0.909	0.867
PeFinder	0.867	0.908	0.873	0.927
95% CI for difference	0.002, 0.148*	−0.046, 0.012	−0.044, 0.120	−0.097, -0.025*
Sensitivity (%)				
CNN model	0.950	0.952	0.882	0.847
PeFinder	0.900	0.945	0.912	0.946
95% CI for difference	0.000, 0.125	−0.030, 0.044	−0.100, 0.000	−0.150, -0.050*
Specificity (%)				
CNN model	0.997	0.905	0.998	0.953
PeFinder	0.993	0.929	0.994	0.958
95% CI for difference	−0.001, 0.010	−0.052, 0.002	−0.001, 0.010	−0.027, 0.017
Accuracy (%)				
CNN model	0.995	0.921	0.994	0.921
PeFinder	0.989	0.935	0.991	0.955
95% CI for difference	0.000, 0.012	−0.036, 0.007	−0.003, 0.009	−0.056, -0.013*

* Indicates significant difference between CNN and PeFinder.

samples (sampled with replacement) taken from each validation set, computed the difference, and took the 2.5th and 97.5th percentiles of the differences as the 95% bootstrap CI of the difference. Bootstrap CIs were constructed by using R statistical software (version 3.4.2; R Foundation, Vienna, Austria) (17). If the CI did not include zero, then we concluded there was a significant difference between the performance of PeFinder and the CNN model. In addition to the comparison with PeFinder, we evaluated the performance of the CNN model in classifying subsegmental PE; PeFinder was not designed to classify subsegmental PE. We measured area under the curve, sensitivity, specificity, accuracy, precision, recall, and F1 score for both models in the two validation datasets. The F1 score (also *F* score or *F* measure) can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst value at 0.

Results

The training set of 2512 reports contained 1357 (54%) PE-positive reports, of which 1190 (88%) were acute and 242 (18%) were subsegmental. The calibration set of 1000 reports contained 50 (5%) PE-positive reports, of which 42 (84%) were acute and eight (16%) were subsegmental. The internal validation of 1000 reports contained 40 (4%) PE-positive reports, of which 34 (85%) were acute and 11 (28%) were subsegmental. The external validation set of 859 reports contained 293 (34%) PE-positive reports, of which 261 (89%) were acute. The external validation set did not distinguish subsegmental location.

All performance results for the CNN model and PeFinder are listed in the Table. In 990 of 1000 cases in the internal validation set, the model correctly predicted all three classification dimensions (accuracy of 99%). In the internal validation set, the CNN model had a statistically significant larger F1 score (0.938) than did PeFinder (0.867) when classifying PE as positive or negative, but no significant difference was found in sensitivity, specificity, or accuracy (Fig 2). In addition, no statistical difference was found in performance when classifying acute or chronic PE in the internal validation set. In the external validation set, on which PeFinder was trained, the 95% CI does not include the value of zero effect. Thus, no significant difference was found between the performance of the CNN model and PeFinder in classifying PE as positive or negative, but PeFinder had significantly larger F1 score, sensitivity, and accuracy when classifying PE as acute or chronic.

The percentage agreement for the internal validation set was 0.99 for all three tasks. The interrater reliability κ scores were 0.96, 0.97, and 0.66 for present, acute, and subsegmental-only PE, respectively. Examples from the internal validation test set from the qualitative visualization method are shown in Figure 3.

Discussion

Our study examined the use of CNNs in classifying free-text medical imaging reports based on the presence, chronicity, and location of PE and compared CNN performance on reports from two different institutions to the best-available NLP model for this task, PeFinder. We found that our trained

Figure 2

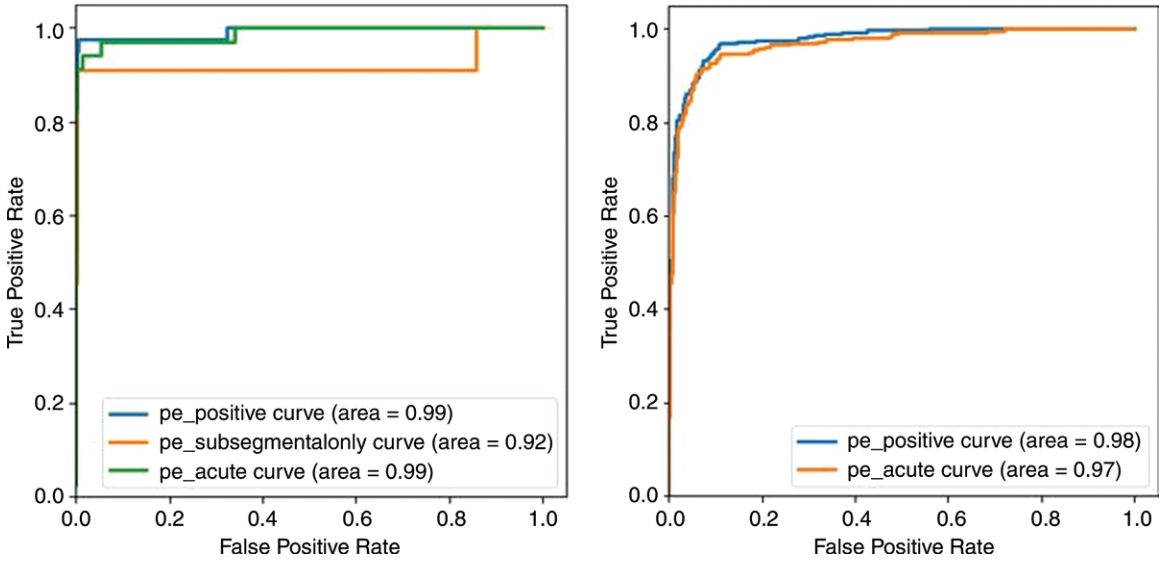

Figure 2: Graphs show receiver operating characteristic curves for each labeling task by using CNN classifier in internal validation set (left) and external validation set (right).

Figure 3

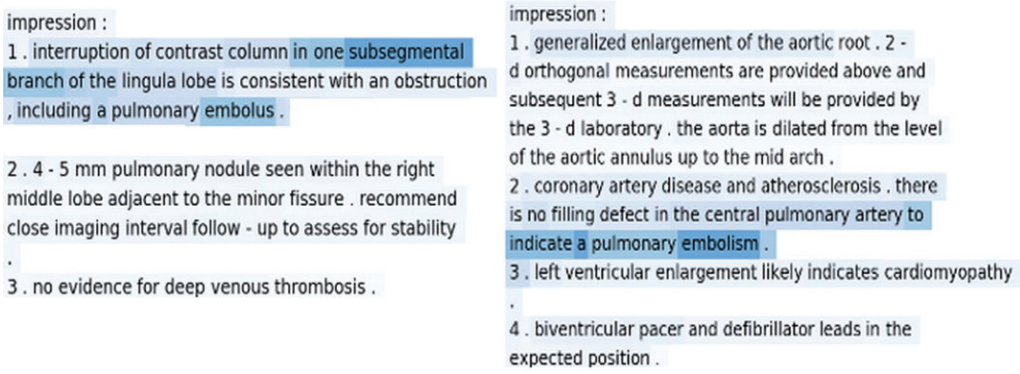

Figure 3: Image shows visualization analysis of input for two examples that CNN classifier predicted correctly (left panel) and incorrectly (right panel) from the internal validation set. Based on intensity of color highlighting, there is little importance placed on the long document with exception of relevant phrases. Result in left panel is example of a report that is PE positive, acute, and subsegmental only; in this case, mention of pulmonary embolus is separated from "one subsegmental branch" and yet model was able to correctly identify the label. Result on right panel is text of report impression that the model predicted was PE positive, but the ground truth was PE negative. Distance between words "no" and "pulmonary embolism" may have contributed to classification error.

CNN model can accurately predict the presence and characteristics of PE from unstructured text in CT reports with an accuracy of 99% and an area under the curve value of 0.97. Compared with PeFinder, the CNN model also performed significantly better in annotation task performance in the intrainstitutional validation set and achieved equivalent performance on the interinstitutional validation reports obtained from the PeFinder institution.

Computational NLP methods make it possible to derive useful structured data from large repositories of narrative medical data (18,19). Chapman et al (8) developed the application PeFinder, based on an extension of NegEx, to classify CT reports for the presence and characteristics of PE. Similarly, Yu et al (11) used the NLP system Narrative Information Linear Extraction (NILE), which combined linguistic and machine learning approaches to improve identification of PE location. Although accurate, these approaches to NLP rely on time-consuming and task-specific

feature engineering such as developing term definitions, syntax annotation, and laborious coding for specific terms. These resources require a great deal of development effort and represent, in certain cases, decades of prior work. In contrast, the CNN model used in this experiment was developed without first defining terms, phrases, or semantic input for the task of identifying PE. Instead, the CNN model was simply trained on a small proportion of labeled reports with a clearly defined input (impression free text) and output (annotation labels) and achieved superior performance levels. Because CNN models have a low burden of development to extract features from medical imaging text, they may be more efficient and scalable for other clinical NLP processing tasks.

The accuracy measures of the CNN model are comparable to or superior to those of other published NLP studies. Ten misclassifications occurred among the 1000 test cases. Ordinarily, neural network classification errors are difficult to explore because the models are considered a black box; however, the visualization methods we used enable an understanding of the source of errors. The words and phrases (including syntax and negation) manifested organically, requiring no engineering or user input, and the visualization technique confirms that the expected words and phrases important to the output classification task were used by the model. The use of CNN models in this way lowers the barrier for NLP development and introduces profound flexibility to the task. For example, the decision to add the subsegmental category was included in our study as a test of this concept of how to add a new labeling task beyond the categories the PeFinder was engineered to annotate. As with the other labeling steps for "PE present" or "acute PE," only document-level annotations were needed to train the model for the new labeling task. In other words, to add a new classification task to the model, the only requirement is the addition of training data via document-level labels, rather than additional coding for new words, phrases, syntax, et cetera.

To explore how the CNN model would generalize to chest CT reports from other institutions, we benchmarked our model against the performance of PeFinder. To design a valid comparison, the external validation set was the same data set used in the initial PeFinder publication (8). Although performance of the CNN model decreased in the external validation set, we found no statistically significant difference in performance between the CNN model and PeFinder. For determining PE chronicity, performance was equivalent in the internal validation set, whereas PeFinder performed significantly better in the external validation set. Overall, the results indicate that the CNN model generalized well to reports that were not included in the training data. In fact, we observed that both models generalized to reports from another institution not used for training. The performance of PeFinder in the internal validation set was also excellent. Applying the CNN model to reports from a larger range of institutions could help confirm the generalizability of this model and approach.

Use of a CNN model for information extraction from natural language reports may yield insights at a scale that would be difficult to obtain with manual annotation. The use of neural networks for report labeling could demonstrate value in research applications, such as cohort generation in precision health research and eligibility screening in clinical trials. Clinical applications might include the use of diagnostic yield as a metric in radiology clinical decision support to manage imaging utilization. For example, the rate of negative studies in similar patients could serve as an indirect marker of utilization appropriateness to guide clinical decision making (20–22). More recently, the surge in computer vision research in radiology has led to a demand for labeled medical images to provide cohorts of training and test data; CNN models such as the one demonstrated here could provide structured labels automatically for computer vision experiments, which require a large corpus of annotated imaging data (23).

Our study had limitations. The machine learning classifier did not achieve perfect accuracy, yet the accuracy measures are comparable or superior to other published NLP studies, as stated. We included only radiology reports from two large tertiary care centers and the performance obtained may be biased. In addition, no structured reports were included in our study, despite that they are increasingly used in many practices. However, information extraction from structured reporting is not felt to pose as great of a challenge compared with free text. We did not explore other CNN variants such as recursive neural networks and subtypes such as long short-term memory, which are designed for text analysis in complex applications. Future work may explore these and other evolving deep learning techniques to classify large volumes of text (24).

In conclusion, we report the results of a CNN model that can accurately classify free-text chest CT reports based on predefined criteria demonstrating both intra- and interinstitutional fidelity when compared with more laborious NLP methods; these results suggest feasibility of CNNs in the automated classification of clinical imaging reports. These machine learning techniques may have important clinical and research applications, even when applied to data from multiple institutions.

Disclosures of Conflicts of Interest: M.C.C. disclosed no relevant relationships. **R.L.B.** disclosed no relevant relationships. **L.Y.** disclosed no relevant relationships. **N.M.** disclosed no relevant relationships. **B.E.C.** disclosed no relevant relationships. **D.B.L.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: reports grants/grants pending to author's institution by Philips and Siemens; receives royalties for intellectual property license from Bayer. Other relationships: disclosed no relevant relationships. **C.P.L.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is founder, shareholder, and board member of Montage Health Solutions; received consulting fees and travel reimbursement from Montage Health Solutions. Other relationships: disclosed no relevant relationships. **T.J.A.** disclosed no relevant relationships. **M.P.L.** Activities related to the present article: reports grants from Philips and the Stanford Child Health Research Institute. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships.

References

1. Dreyer KJ, Kalra MK, Maher MM, et al. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology 2005;234(2):323–329.
2. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161–174.
3. Hassanpour S, Langlotz CP, Amrhein TJ, Befera NT, Lungren MP. Performance of a machine learning classifier of knee MRI reports in two large academic radiology practices: a tool to estimate diagnostic yield. AJR Am J Roentgenol 2017;208(4):750–753.
4. Hripcsak G, Austin JH, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 2002;224(1):157–163.
5. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444.
6. Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging 2017;30(4):427–441.
7. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in x-ray images. Med Image Anal 2017;36:41–51.
8. Chapman BE, Lee S, Kang HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform 2011;44(5):728–737.
9. Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Ther 2015;8:2015–2022.
10. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017;284(2):574–582.
11. Yu S, Kumamaru KK, George E, et al. Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing. J Biomed Inform 2014;52:386–393.
12. Cai T, Giannopoulos AA, Yu S, et al. Natural language processing technologies in radiology research and clinical applications. RadioGraphics 2016;36(1):176–191.
13. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. Stanford, Calif: Stanford University, 2014. https://nlp.stanford.edu/pubs/glove.pdf.
14. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, October 25–29, 2014. Stroudsburg, Pa: Association for Computational Linguistics, 2014; 1746–1751.
15. Mikolov T, Sutskever I, Chen K, Corrado C, Dean J. Efficient Estimation of Word Representations in Vector Space. In: NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, December 5–10, 2013. Red Hook, NY: Curran Associates, 2013; 3111–3119.
16. Arras L, Horn F, Montavon G, Müller KR, Samek W. Explaining Predictions of Non-Linear Classifiers in NLP. In: Proceedings of the 1st Workshop on Representation Learning for NLP, Berlin, Germany, August 11, 2016. Stroudsburg, Pa: Association for Computational Linguistics, 2016; 1–7.
17. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2017.
18. Gallego B, Dunn AG, Coiera E. Role of electronic health records in comparative effectiveness research. J Comp Eff Res 2013;2(6):529–532.
19. Leeper NJ, Bauer-Mehren A, Iyer SV, Lependu P, Olson C, Shah NH. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PLoS One 2013;8(5):e63499.
20. Lungren MP, Amrhein TJ, Paxton BE, et al. Physician self-referral: frequency of negative findings at MR imaging of the knee as a marker of appropriate utilization. Radiology 2013;269(3):810–815.
21. Amrhein TJ, Lungren MP, Paxton BE, et al. Journal Club: Shoulder MRI utilization: relationship of physician MRI equipment ownership to negative study frequency. AJR Am J Roentgenol 2013;201(3):605–610.
22. Paxton BE, Lungren MP, Srinivasan RC, et al. Physician self-referral of lumbar spine MRI with comparative analysis of negative study rates as a marker of utilization appropriateness. AJR Am J Roentgenol 2012;198(6):1375–1379.
23. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–248.
24. Luo Y. Recurrent neural networks for classifying relations in clinical notes. J Biomed Inform 2017;72:85–95.