Journal of Digital Imaging (2023) 36:105–113
https://doi.org/10.1007/s10278-022-00712-w

ORIGINAL PAPER

Natural Language Processing Model for Identifying Critical Findings—A Multi-Institutional Study

Imon Banerjee[1,2] · Melissa A. Davis[3] · Brianna L. Vey[3] · Sina Mazaheri[3] · Fiza Khan[3] · Vaz Zavaletta[3] · Roger Gerard[3] · Judy Wawira Gichoya[3] · Bhavik Patel[1,2]

Received: 12 January 2022 / Revised: 2 September 2022 / Accepted: 3 October 2022 / Published online: 7 November 2022

Abstract

Improving detection and follow-up of recommendations made in radiology reports is a critical unmet need. The long and unstructured nature of radiology reports limits the ability of clinicians to assimilate the full report and identify all the pertinent information for prioritizing the critical cases. We developed an automated NLP pipeline using a transformer-based ClinicalBERT^{++} model which was fine-tuned on 3 M radiology reports and compared against the traditional BERT model. We validated the models on both internal hold-out ED cases from EUH as well as external cases from Mayo Clinic. We also evaluated the model by combining different sections of the radiology reports. On the internal test set of 3819 reports, the ClinicalBERT^{++} model achieved 0.96 f1-score while the BERT also achieved the same performance using the reason for exam and impression sections. However, ClinicalBERT^{++} outperformed BERT on the external test dataset of 2039 reports and achieved the highest performance for classifying critical finding reports (0.81 precision and 0.54 recall). The ClinicalBERT^{++} model has been successfully applied to large-scale radiology reports from 5 different sites. Automated NLP system that can analyze free-text radiology reports, along with the reason for the exam, to identify critical radiology findings and recommendations could enable automated alert notifications to clinicians about the need for clinical follow-up. The clinical significance of our proposed model is that it could be used as an additional layer of safeguard to clinical practice and reduce the chance of important findings reported in a radiology report is not overlooked by clinicians as well as provide a way to retrospectively track large hospital databases for evaluating the documentation of the critical findings.

Introduction

Radiology reports not only document the radiologist's observations and diagnosis of disease observed on images, but importantly, they also convey a description of critical and incidental findings, as well as recommendations for follow-up of those findings [1]. A critical finding is defined as a finding in imaging exam that requires immediate or urgent communication with the provider (person who placed the order) since these findings reflect conditions that are life-threatening (e.g., tension pneumothorax) or conditions that require an immediate change in the patient management (e.g., retained surgical objects) [2]. It is critical for the physicians to be aware of and keep track of these findings in order to render timely medical care. However, in the current high-volume environment of clinical care with rising physician burn-out, there can be a tendency for clinicians to focus on the urgent current clinical issues first and non-urgent follow-up on other tangential issues later [3]. Consequently, important incidental findings are often not followed up by the physicians. In fact, one study found that in 51% of cases, physicians failed to obtain the recommended follow-up exams. Moreover, litigation with settlements of up to $2.5 M from physicians and hospitals related to the lack of following recommendations for critical or incidental findings have been reported [4].

Identifying and tracking critical findings in radiology reports provides an efficient way for the automated generation of alerts for recommendations of further follow-up. Theoretically, extraction of the radiological findings should be a straightforward task since the American College of Radiology

✉ Imon Banerjee
Banerjee.Imon@mayo.edu

[1] Department of Radiology, Mayo Clinic, 5777 E Mayo Blvd, Phoenix, AZ 85054, USA

[2] Arizona State University, SCAI, 6161 E Mayo Blvd, Phoenix, AZ 85054, USA

[3] School of Medicine, Emory University, 1364 Clifton Road NE, 30322 Atlanta, USA

(ACR) outlines a structured format of the radiology report for the communication of diagnostic imaging findings which states that imaging findings are contained in the body of the report, while specific diagnosis should be given in the separate "impression." However, despite these guidelines, radiology reporting still includes extensive variations—starting from the usage of varying terms for describing findings—to the section within which the findings are documented. Thus, given the linguistic and structural complexity of the radiologic documentation, identification of critical findings becomes a challenging problem and thus, often, the critical imaging findings are missed in this universal form of communication between radiologists and referring physicians. Although, it could be well suited for an informatics solution via computerized text analysis which can incorporate semantic language space to track the synonyms and acronyms of the findings and allow the combinations of multiple sections of the radiology reports. Such systems not only allow to improve the communication between the care team but also provide a way to retrospectively track large hospital databases for evaluating the documentation of the critical findings and identifying "missed" cases which can ultimately contribute to healthcare quality assessment and epidemiological studies.

Prior work in this area has used simplistic rule-based methods that do not perform sufficiently accurately [5] and often failed to present the generalization of the data from centers other than those used to develop the system [6–8]. A major limitation of natural language processing (NLP) methods developed to date is limited generalizability; their performance varies depending on the data used to train the models, and they often do not perform well on data other than those used to train the model [8]. To address the potential limitation of generalizability, one needs to collect radiology reports from multiple different sites to obtain dataset that is more representative of the populations on which the model will be applied and report the performance on both internal and external datasets. However, there are major barriers to sharing data among hospitals due to privacy restrictions which limits the evaluation of the generalizability of the existing NLP systems.

Our objective is to develop an automated NLP system that can analyze free-text radiology reports, along with the reason for the exam, to identify critical findings from large-scale hospital databases and raise alerts for the missed cases. Such model needs to be trained and validated with a large-scale multi-modal dataset to capture different types of critical findings (e.g., neurologic—hemorrhage, GU—Ectopic pregnancy, cardiac—acute myocardial infraction) from radiologic reporting of various modalities (MR, CT, Ultrasound). In this study, we proposed a transformer based [9] ClinicalBERT^{++} model by fine-tuning the ClinicalBERT model on 3 M radiology reports and validated the performance using both internal cases from Emory University Hospital (EUH) and external ED cases from Mayo Clinic. We performed a detailed evaluation of the model using different settings and presented error analysis for model explainability using interactive model visualization of LIME.

Methods

Figure 1 presents the overall pipeline and subsections below and describes each processing block.

Dataset

Ethical review was obtained from the Emory Institutional Review Board (IRB) with permission to extract radiology reports for patients who had an emergency room visit in 2019 at Emory University Hospital (EUH). Retrospective

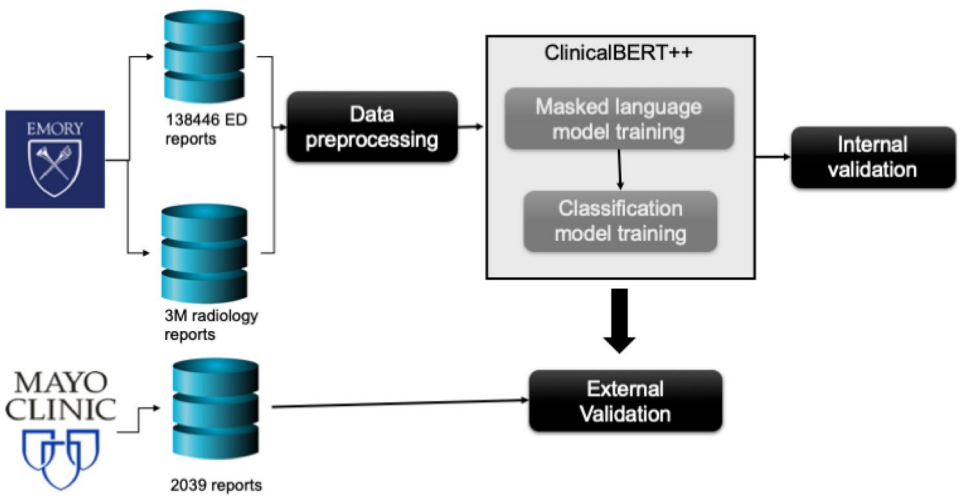


Fig. 1 The NLP Pipeline for developing the system and validation strategy. Showing both the internal and external validation strategy

manual labeling of 138,446 radiology reports was performed by 5 radiology residents (1–3 years of experience) under the supervision of attendings (5–10 years). An annotation key (see Appendix) developed by ACR's Actionable Reporting Work Group was used to analyze the reports. Among the 5 radiology residents, the agreement score (Fleiss kappa) was 0.81 on 100 common cases which shows that there is a substantial agreement between the annotators. In addition to the internal dataset, we also obtained the approval from Mayo Clinic IRB and retrieved the radiology reports from patients who had an emergency room visit in 2019. We used the Mayo Clinic data only for external validation and manually collected labels for randomly selected 2039 cases from reading of two attendings (Cohen kappa 0.76). Similar to EUH, the reports were annotated retrospectively by analyzing the radiology report text. Figure 2 shows the distribution of data in each category of EUH and Mayo Clinic reports, including the acquisition modalities.

Report Pre-Processing

We extended our NLP methods to parse the clinical history, imaging protocol, findings, and impression sections of the radiology reports using section segmentation based on the header. In order to generalize the section segmentation across multiple institutions, we extracted all the variations of the headers using a similar word list generated by the Word-2Vec language model [10] trained on 3 M radiology reports from EUH. Such non-contextual language model generates a similar word list only by reflecting co-occurrence statistics which is sufficient for capturing header variations between reports given that such words appear in similar contexts. We computed the similar word list for each header by intersecting the list of other headers. For example, generated similar wordlists for the "clinical history" section include "indications," "history," "patient history," "reason for exam," and "reason for order." After the section segmentation, we used the terms and their synonyms from the RadLex ontology [11] to normalize words across modalities by mapping to the root terms. All the communication statements between the radiology and care team were removed from the radiology reports before parsing to reduce the information leakage in the model training.

Model Training and Validation

After the section segmentation, we developed and tested classifier models by combining different sections of the radiology reports. The EUH data were split randomly in training validation and test sets in a 60:20:20 ratio and the majority class (non-acute) was down-sampled for training. We used Mayo Clinic data for external validation only. For classifying the reports, we adopted two distinct transformer models—BERT and ClinicalBERT^{++}. BERT [12] architecture uses bi-directional transformers and generates contextualized word representations by training a masked language model. It has been proven to be one of the most powerful natural language processing models to date and had improved the state-of-art performance on at least 11 tasks when it was published. BERT is pre-trained on two unsupervised learning tasks of English news text, masked language model, and next sentence prediction. In contrast, ClinicalBERT [13], was initially trained on PubMed abstracts (PubMed), Central full-text articles (PMC), and all MIMIC notes (implementation available on [2]) and expected to capture the clinical language space. In order to capture the radiology language space, we use masked language modeling tasks to fine-tune pre-trained ClinicalBERT model using masked language model on 3 M radiology reports (clinical history, finding, and impression section) from EUH and developed a new

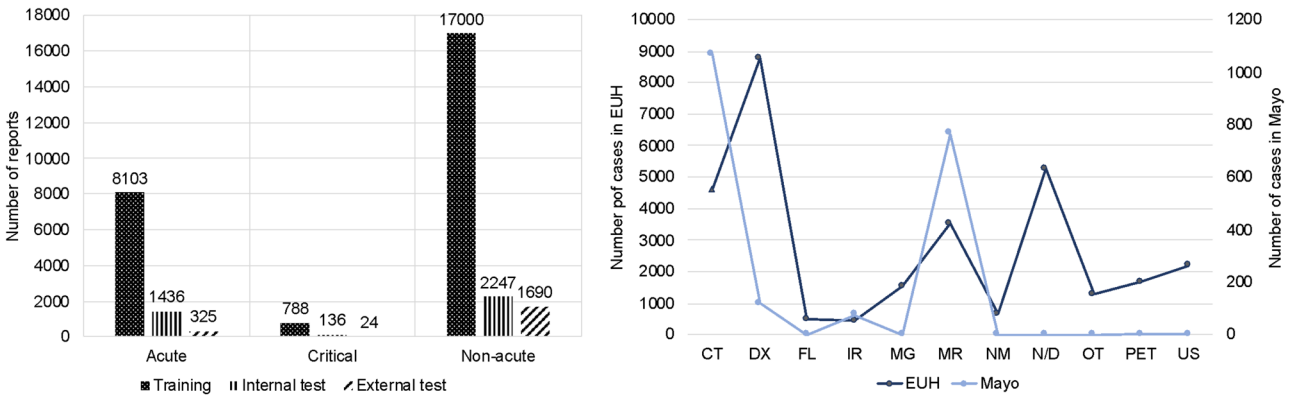


Fig. 2 Distribution of data—EUH (training and internal test) and Mayo Clinic (external test)—based on the categories (left) and based on the modalities (right)

version so the model would learn specific domain knowledge regarding radiology. The models were fine-tuned on a high-performance computing cluster with PyTorch, Pandas, and Numpy libraries with modified source code from HuggingFace Transformers. Models were trained on NVIDIA Tesla V100 GPUs with 32 GB of memory.

After fine-tuning the language space, the ClinicalBERT^{++} model was trained for the critical finding classification task. Hyperparameter configurations were finalized by grid search on the validation data, with a batch size of 20 for smaller models, 10 epochs, dropout of 0.1, weight decay of 0.1, and the AdamW optimizer ($\varepsilon = 10–8$, $\beta 1 = 0.9$, $\beta 2 = 0.999$). During the evaluation, beam search with 10 beams and forbid trigram repetition was used. The model was then applied to a cohort of reports generated in the care of patients in the Emergency Department for each of the 5 hospitals. A report acuity score was assigned as the proportion of reports classified as acute and critical to those that were flagged as normal. After the training completion, we externally tested the model on the Mayo Clinic data without any alteration. The Mayo Clinic reports were also pre-processed using the same codebase developed for EUH.

Results

Figure 3 presents the word distribution statistics across EUH and Mayo Clinic for each class label where tags are top 200 frequent single words or bigrams, and the frequency of each tag appearance is shown with font size. As seen from the representation, common critical findings in EUH are usually trauma related while in Mayo, there are rare trauma-related words that appear in the frequent word list. Given the differences in practice setting types (e.g., trauma center vs. oncology center), the generalization is even more challenging and resultant a simultaneous effect in language space. However, we purposefully fine-tuned the ClinicalBERT weights on the 3 M generic radiology reports to account for such variations.

Table 1 shows the performance of both BERT and ClinicalBERT^{++} on the hold-out internal test set from EUH and the external set from Mayo Clinic using only the finding or the impression section. The ClinicalBERT^{++} performed optimally using the impression section rather than findings. This could be due to the summarized information representation in the impression section and the fact that transformer models are sensitive to text length as the model will consume a maximum of 512 tokens and truncate anything beyond the length. ClinicalBERT^{++} model achieved 0.87 overall f1-score and 0.73 f1-score for the Critical finding category. However, the BERT model on the internal hold-out test data outperformed ClinicalBERT^{++} with 0.95 overall f1-score and 0.82 f1-score for the Critical finding. On the other hand, it failed to retain the similar performance on the for critical finding reports from the external test data and achieved only 0.44 f1-score while ClinicalBERT^{++} achieved 0.65 f1-score. The overall weighted f1-score stays similar between internal and external test datasets.

In order to reduce the false positive and negative for critical findings, we combined the reason for the exam along with the impression section and evaluated both the models in Table 2. Given the chance to analyze the reason for ordering the exam, the performance of both the models improved up to 0.96 overall f1-score on the same internal test data in this setting. ClinicalBERT^{++} outperformed BERT on the external test dataset and achieved the highest performance for classifying critical finding reports (0.81 precision and 0.54 recall) when the BERT model achieved 0.23

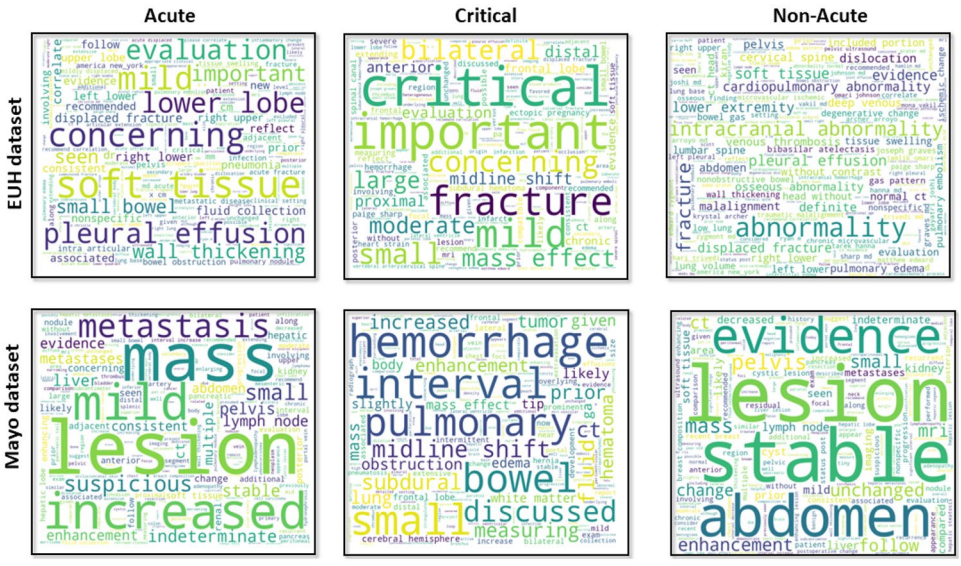


Fig. 3 Word-cloud representation of EUH and Mayo dataset—stratified by class labels

Table 1 Performance of the models using only finding or only impression section. Bold text represents optimal performance on the specific dataset

Labels	Model	Finding section			Impression section			Support
		Precision	Recall	F1-score	Precision	Recall	F1-score	
Internal hold-out test set								
Acute	BERT	0.93	0.95	0.94	0.91	0.72	0.80	1436
Critical		0.82	0.82	0.82	0.76	0.67	0.71	136
Non-acute		0.98	0.96	0.97	0.84	0.96	0.90	2247
Overall (weighted)		**0.95**	**0.95**	**0.95**	0.87	0.86	0.86	3819
Acute	ClinicalBERT^{++}	0.91	0.72	0.80	0.92	0.75	0.83	1436
Critical		0.76	0.67	0.71	0.78	0.68	0.73	136
Non-acute		0.84	0.96	0.90	0.86	0.96	0.91	2247
Overall (weighted)		0.87	0.86	0.86	**0.88**	**0.87**	**0.87**	3819
External test								
Acute	BERT	0.06	0.15	0.09	0.79	0.81	0.80	325
Critical		0.20	0.33	0.25	0.67	0.33	0.44	24
Non-acute		0.76	0.53	0.63	0.96	0.96	0.96	1690
Overall (weighted)		0.64	0.47	0.54	**0.93**	**0.93**	**0.93**	2039
Acute	ClinicalBERT^{++}	0.06	0.16	0.09	0.64	0.55	0.60	325
Critical		0.26	0.38	0.31	0.81	0.54	0.65	24
Non-acute		0.77	0.55	0.64	0.92	0.94	0.93	1690
Overall (weighted)		0.65	0.49	0.55	0.87	0.88	0.87	2039

precision and 0.33 recall. This drop in the performance is primarily due to the linguistic variations between the internal and external datasets as well as the class-imbalanced during the model training. However, due to the representation of the realistic performance, we did not try to artificially balance the dataset. The model primarily contains the false negative cases and only 15% false positive for critical finding.

For ClinicalBERT^{++}, we explored the reason for the false negative cases for critical findings using LIME [14] which can generate model-agnostic visual explanations of machine learning models by weighting the importance of each word. Figures 3 and 4 show false negative examples of the ClinicalBERT^{++} model using combined text from clinical history and impression section where the weights are represented as class-wise colormap. Figure 4 represents a case from the external Mayo Clinic dataset where the model incorrectly predicted a critical case as non-acute. The model correctly identified bowel "perforation" as a highly weighted

Table 2 Performance of the models using both clinical history and impression section. Bold text represents optimal performance on the specific dataset. Shows the *p*-value for model comparison

Labels	**Model**	**History and impression section**				**Support**
		Precision	Recall	F1-score	***p*-value**	
Acute	BERT	0.94	0.95	0.95	$p > 0.5$	1436
Critical		0.82	0.86	0.83		136
Non-acute		0.97	0.97	0.97		2247
Overall (weighted)		**0.96**	**0.96**	**0.96**		3819
Acute	ClinicalBERT^{++}	0.93	0.96	0.95	ref	1436
Critical		0.92	0.84	0.88		136
Non-acute		0.98	0.96	0.97		2247
Overall (weighted)		**0.96**	**0.96**	**0.96**		3819
	External test					
Acute	BERT	0.31	0.89	0.45	$p < 0.01$	325
Critical		0.23	0.33	0.27		24
Non-acute		0.97	0.61	0.75		1690
Overall (weighted)		0.85	0.65	0.69		2039
Acute	ClinicalBERT^{++}	0.64	0.55	0.60	ref	325
Critical		0.81	0.54	0.65		24
Non-acute		0.92	0.94	0.93		1690
Overall (weighted)		**0.87**	**0.88**	**0.87**		2039

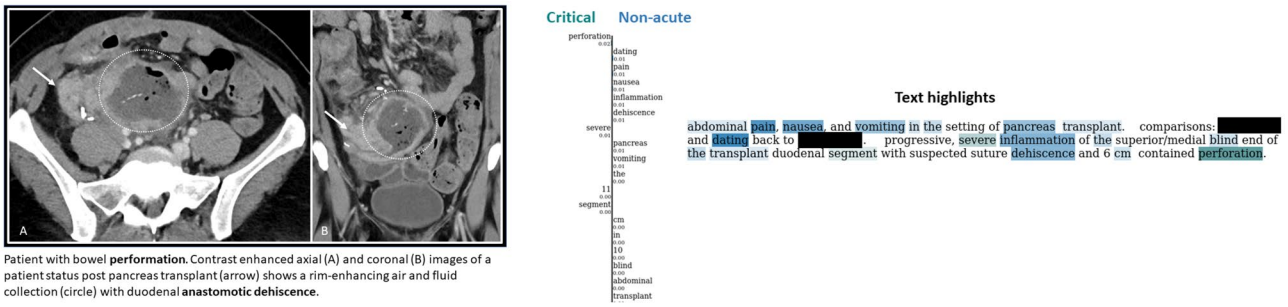
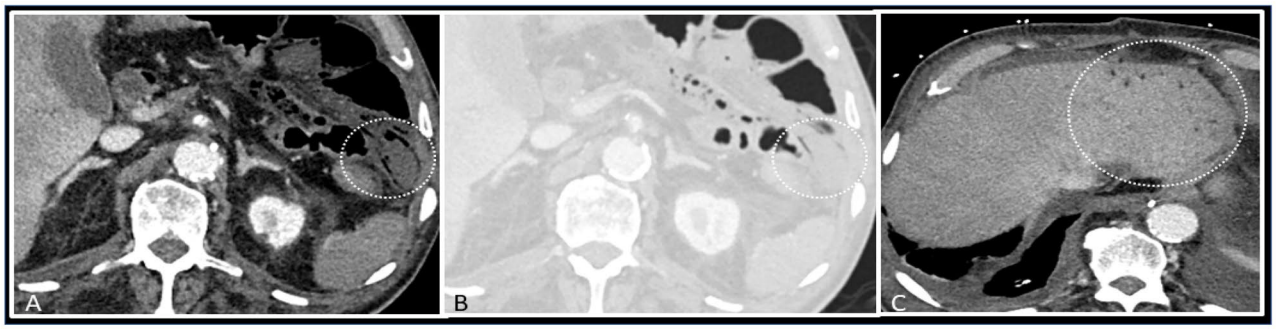


Fig. 4 A critical finding case from the external validation dataset classified by the model as non-acute—(left) we highlight the imaging finding and (right) we present the LIME text highlight for both the wrongly predicted class non-acute and original class critical finding. Darker color represents higher weights. Dates are blacked out for preserving patient privacy

word for the report to be critical but given the pancreas transplant history, the model assigned more weight to the non-acute words, like "inflammation."

Figure 5 represents another case where the model is predicted as acute while the case is actually critical. Similar to the previous example, "ischemia" is identified as an important word for classifying as critical but "pneumatosis" and "bowel" issues pull the classification decision towards acute. Based on these explanations, we can note the fact that the current model has a tendency to classify critical findings



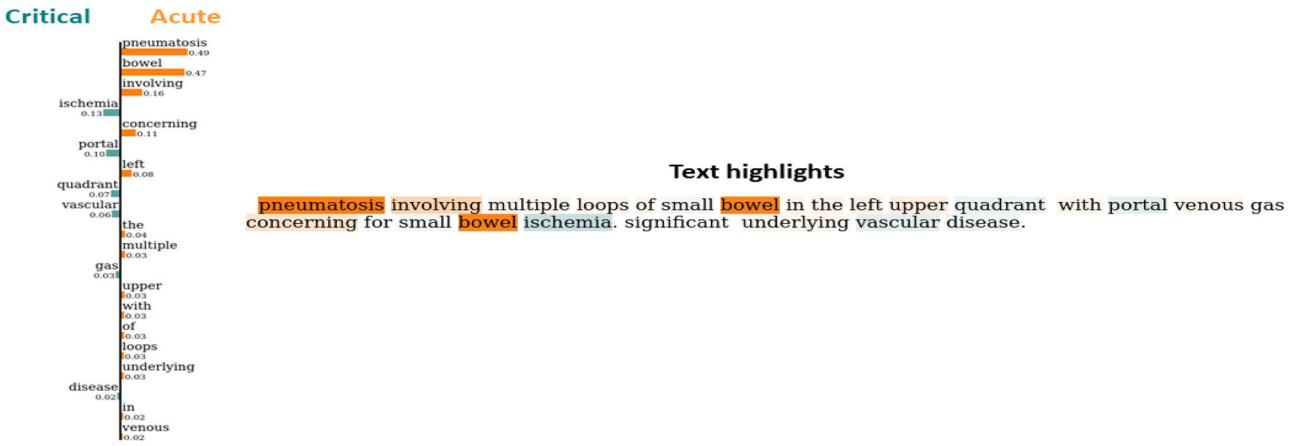


Fig. 5 A critical finding case from the external validation dataset classified by the model as Acute—(left) we highlight the imaging finding and (right) we present the LIME text highlight for both the wrongly predicted class acute and original class critical finding. Darker color represents higher weights

Table 3 Number of acute, non-acute, and critical cases (% of cases) in 5 distinct hospitals—annotated by the ClinicalBERT^{++} model

	Total radiology reports	Critical	Acute	Non-acute	*p-value
Hospital 1	166,789	2,635 (1.58%)	45,573 (27.32%)	118,581 (71.10%)	***
Hospital 2	65,035	835 (1.28%)	9,263 (14.24%)	54,937 (84.47%)	$p<0.0001$
Hospital 3	44,703	512 (1.15%)	6,475 (14.48%)	37,716 (84.37%)	$p<0.0001$
Hospital 4	32,684	408 (1.25%)	4,609 (14.10%)	27,667 (84.65%)	$p<0.0001$
Hospital 5	17,345	218 (1.26%)	2,415 (13.92%)	14,712 (84.82%)	$p<0.0001$

as acute of reports mention multiple findings within close vicinity of the critical finding texts.

Summary on Running with Multiple Centers

To understand the scalability of the ClinicalBERT^{++} model, we applied the trained model to all the ED cases from four different hospital systems in the Emory system. Hospital 1 is a designated Safety Net Hospital (SNH) and Level 1 Trauma Center, hospital 2 is on the same campus as the School of Medicine, the undergraduate and other graduate schools, hospital 3 is in the urban center, and hospitals 4 and 5 are community-based hospitals in the more suburban areas of the city. Hospitals 1–3 are the primary teaching hospitals. Hospital 1 has the largest number of radiology exams performed in the emergency room ($n=166{,}789$) compared to hospitals 2–5 (Table 3). We found that report acuity did vary from site to site and hospital 1 which is a SNH, had a statistically significant higher proportion of Critical and Acute reports (28.9%), compared to all the other hospitals. Hospital 1 is also the only level 1 trauma center in the system so the finding of the highest report acuity is consistent with the expectation.

*Acute and critical cases/total reports in hospital 1 were compared to the proportion in each of the other hospitals.

Discussion

Radiologic findings often guide the next steps for patients, whether it be transferring to an operating room or an intensive care unit, placement in a floor bed, or discharge. Acuity identified in the radiology reports can also be leveraged as a triage process for patients immediately within their care pathway. However, a large portion of the radiology reports exists in an unstructured format making it difficult to extract insights rapidly across multiple reports from multiple modalities despite expanding use of imaging for diagnosis and patient triage. However, there are various open-source libraries and NLP tools available that facilitate the automated extraction of information from the unstructured and semi-structured reports to the benefit of the data mining of radiology. Such research endeavors can also facilitate epidemiological studies involving large amounts of human language.

We develop an automated NLP system that can rapidly parse a large set of free-text radiology reports from multiple domains to identify critical and acute radiology findings to enable and alert notifications to clinicians about the need for clinical follow-up as well as generate a curated large-scale database with critical and acute finding labels. Most of the existing systems are rule-based and thus, are non-trivial to generalize outside the organization of training due to variations in the language and templates. Our experimentation showed that adding the clinical history section allow the model to understand the context of the study and such case context improved the model performance for determining the critical vs acute classification. We believe the model is simulating clinical practice by learning how humans contextualize results in determining the report label. For example, pneumoperitoneum in the post-operative period without clinical history or reason for exam information may be misclassified as a critical finding. Such flagging may inadvertently create additional work for the radiologists having to correct it, or worse if not corrected, could potentially have downstream consequences as a result. With the context of a recent surgery, the NLP accurately classifies the pneumoperitoneum as an expected finding.

Our approach addresses the two major limitations of efficiency in multiple radiology domains and generalizability that have hindered the prior approaches to this problem, as our system adopted transformer architecture fine-tuned with a large number of radiology cases to create a radiology-specific language space. It learnt to recognize and unify the diverse ways that the same concepts are expressed in narrative reports and provided high accuracy for detecting critical findings from different types of radiology reports performed in ED. We compared ClinicalBERT^{++} model against the state-of-the-art BERT model which was primarily trained on news articles and achieved performance improvement for the critical finding class. The proposed NLP model presents high accuracy for classifying acute, non-acute, and critical classes across multiple sites, even when the language space is significantly different (Fig. 2). The NLP model has been successfully applied to large-scale radiology reports from 5 different sites. Even though the ground truth labels were not available for those cases to measure the performance, the findings are consistent with the human expectation.

The model failed for cases where multiple findings are represented in a non-structured way within the radiology impression section (Figs. 4 and 5). Even though the transformer model trained in unsupervised way using a large dataset can handle significant variations in terminology, the structural variations of reporting (e.g., multiple findings in the same sentence) still affect the model performance. In order to obtain optimal performance, standardized radiology reporting may play a critical role. To address the challenge of generalizability, future work will use federated learning to leverage reports from multiple institutions thereby facilitating training on more representative diverse data. A robust automated system that can incorporate NLP models can provide safeguards to clinical practice and reduce the chance of important observations being overlooked by clinicians. Rather than the automated insertion of the model finding in the reporting, we hypothesized that our system could ultimately be integrated into the radiology workflow as an alerting system from both physicians and patients for appropriate follow-up by interfacing it to the HL7 feed.

The ClinicalBERT^{++} model will be freely available with open-source licensing in for community development. Such radiology-specific language models trained on large-scale multi-modal data could play an important role in developing specific NLP solutions for different radiology use cases. The clinical significance of the proposed model is that it could be used as an additional automated layer of safeguard to clinical practice and reduce the chance of important findings reported in a radiology report not be overlooked by clinicians.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-022-00712-w.

Author Contribution All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Imon Banerjee, Melissa A. Davis, Brianna L. Vey, Sina Mazaheri, Fiza Khan, Vaz Zavaletta, and Roger Gerard. The first draft of the manuscript was written by Imon Banerjee and all authors commented on the manuscript. All authors read and approved the final manuscript.

Funding The manuscript is partially supported by NIH/NHLBI, 1R01HL155410-01A1: applied deep learning in pulmonary embolism outcome prediction using imaging and clinical data: a multicenter study.

Declarations

Ethics Approval This is an observational study. The Emory University and Mayo Clinic Research Ethics Committee have confirmed that no ethical approval is required.

Consent to Participate The Emory University and Mayo Clinic Research Ethics Committee has confirmed that no informed consent is required.

Conflict of Interest The authors declare no competing interests.

References

1. P. Cronin and J. V. Rawson, "Review of Research Reporting Guidelines for Radiology Researchers," *Acad Radiol*, vol. 23, no. 5, pp. 537–558, May 2016, https://doi.org/10.1016/j.acra.2016.01.004.
2. T. Mabotuwana, C. S. Hall, and N. Cross, "Framework for Extracting Critical Findings in Radiology Reports," *J Digit Imaging*, vol. 33, no. 4, pp. 988–995, Aug. 2020, https://doi.org/10.1007/s10278-020-00349-7.
3. S. Dutta, W. J. Long, D. F. M. Brown, and A. T. Reisner, "Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings," *Ann Emerg Med*, vol. 62, no. 2, pp. 162–169, Aug. 2013, https://doi.org/10.1016/j.annemergmed.2013.02.001.
4. "Lost Souls®: Detect & Track Radiology Recommendations." [Online]. Available: https://radiology-universe.org/Lost-Souls-Radiology-Recommendations-Tracking-Detection/
5. M. E. Heilbrun, B. E. Chapman, E. Narasimhan, N. Patel, and D. Mowery, "Feasibility of Natural Language Processing–Assisted Auditing of Critical Findings in Chest Radiology," *Journal of the American College of Radiology*, vol. 16, no. 9, pp. 1299–1304, Sep. 2019, https://doi.org/10.1016/j.jacr.2019.05.038.
6. A.-D. Pham *et al.*, "Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings," *BMC Bioinformatics*, vol. 15, p. 266, Aug. 2014, https://doi.org/10.1186/1471-2105-15-266.
7. G. Trivedi, E. R. Dadashzadeh, R. M. Handzel, W. W. Chapman, S. Visweswaran, and H. Hochheiser, "Interactive NLP in Clinical Care: Identifying Incidental Findings in Radiology Reports," *Appl Clin Inform*, vol. 10, no. 4, pp. 655–669, Aug. 2019, https://doi.org/10.1055/s-0039-1695791.
8. M. Yetisgen-Yildiz, M. L. Gunn, F. Xia, and T. H. Payne, "A text processing pipeline to extract recommendations from radiology reports," *J Biomed Inform*, vol. 46, no. 2, pp. 354–362, Apr. 2013, https://doi.org/10.1016/j.jbi.2012.12.005.
9. T. Wolf *et al.*, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," arXiv:1910.03771 *[cs]*, Jul. 2020, Accessed: Dec. 08, 2021. [Online]. Available: http://arxiv.org/abs/1910.03771
10. Y. Goldberg and O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," arXiv:1402.3722 *[cs, stat]*, Feb. 2014, Accessed: Dec. 08, 2021. [Online]. Available: http://arxiv.org/abs/1402.3722
11. J. L. V. Mejino, D. L. Rubin, and J. F. Brinkley, "FMA-RadLex: An application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology," *AMIA Annu Symp Proc*, pp. 465–469, Nov. 2008.
12. A. Adhikari, A. Ram, R. Tang, J. Lin, and D. R. Cheriton, "DocBERT: BERT for Document Classification." [Online]. Available: https://github.com/castorini/hedwig
13. E. Alsentzer *et al.*, "Publicly Available Clinical BERT Embeddings," 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/

14. Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell, "'Why Should You Trust My Explanation?' Understanding Uncertainty in LIME Explanations," arXiv:1904.12991 *[cs, stat]*, Jun. 2019, Accessed: Dec. 08, 2021. [Online]. Available: http://arxiv.org/abs/1904.12991