A Web Application for Adrenal Incidentaloma Identification, Tracking, and Management Using Machine Learning

Wasif Bala[1] Jackson Steinkamp[1] Timothy Feeney[1] Avneesh Gupta[1] Abhinav Sharma[2] Jake Kantrowitz[3] Nicholas Cordella[1] James Moses[1] Frederick Thurston Drake[1]

[1] Boston Medical Center, One Boston Medical Center Pl, Boston, Massachusetts, United States
[2] Department of Family Medicine, Sunnybrook Health Sciences Centre, University of Toronto, Toronto, Ontario, Canada
[3] Department of Internal Medicine, Kent Hospital, Brown University Alpert Medical School, Warwick, Rhode Island, United States

Address for correspondence Wasif Bala, MD, Boston Medical Center, One Boston Medical Center Pl, Boston, MA 02118, United States (e-mail: wasifbala@gmail.com).

Appl Clin Inform 2020;11:606–616.

Abstract

Background Incidental radiographic findings, such as adrenal nodules, are commonly identified in imaging studies and documented in radiology reports. However, patients with such findings frequently do not receive appropriate follow-up, partially due to the lack of tools for the management of such findings and the time required to maintain up-to-date lists. Natural language processing (NLP) is capable of extracting information from free-text clinical documents and could provide the basis for software solutions that do not require changes to clinical workflows.

Objectives In this manuscript we present (1) a machine learning algorithm we trained to identify radiology reports documenting the presence of a newly discovered adrenal incidentaloma, and (2) the web application and results database we developed to manage these clinical findings.

Methods We manually annotated a training corpus of 4,090 radiology reports from across our institution with a binary label indicating whether or not a report contains a *newly discovered* adrenal incidentaloma. We trained a convolutional neural network to perform this text classification task. Over the NLP backbone we built a web application that allows users to coordinate clinical management of adrenal incidentalomas in real time.

Results The annotated dataset included 404 positive (9.9%) and 3,686 (90.1%) negative reports. Our model achieved a sensitivity of 92.9% (95% confidence interval: 80.9–97.5%), a positive predictive value of 83.0% (69.9–91.1)%, a specificity of 97.8% (95.8–98.9)%, and an F1 score of 87.6%. We developed a front-end web application based on the model's output.

Conclusion Developing an NLP-enabled custom web application for tracking and management of high-risk adrenal incidentalomas is feasible in a resource constrained, safety net hospital. Such applications can be used by an institution's quality department or its primary care providers and can easily be generalized to other types of clinical findings.

Keywords
- incidental findings
- software tools
- natural language processing
- electronic medical record

received
April 12, 2020
accepted
July 22, 2020

DOI https://doi.org/10.1055/s-0040-1715892.
ISSN 1869-0327.

Background and Significance

Incidental radiological findings, such as adrenal incidentalomas, are clinically relevant and require longitudinal management. A failure to properly track and manage incidental radiological findings may result in missed diagnoses, including malignancies, that predispose patients to poor health outcomes.[1–5] The rate of adherence to recommendations regarding these incidental findings is low. Although the number of recommended follow-ups for radiological findings continues to increase, only 58% of these recommended follow-ups are completed.[6–9] The care coordination process for incidental findings is particularly complex. The timeline for following up such findings is often months to years away, and there is no structured system to track follow-up recommendations, particularly those found in unstructured free text of a radiology report. If the finding is discovered in the inpatient setting, the hospital clinicians must communicate this information to the outpatient clinicians, who must track these outstanding incidental findings using variable free-text formats within existing electronic medical records (EMRs). At each step of this communication process, information is prone to loss, especially given the increasing clinical complexity of patients in an aging population. In some cases, there is no clear consensus on who bears the primary responsibility for keeping track of the finding (e.g., the patient has no primary care provider [PCP]).

Improved methods for tracking and managing these findings as well as the associated follow-up recommendations are required. Previous work has demonstrated that electronic notification tools can be useful in improving adherence to follow-up imaging test completion,[10,11] highlighting the importance of exploring alternative methods of notifying providers of the need for follow-up. In particular, a centralized system which maintains a database of outstanding incidental findings and allows for multiuser management over time would solve many of these issues, as it would not rely on variable individual workflows. When integrated into an existing EMR, these systems can relay follow-up recommendations to physicians, either directly or with the assistance of a quality improvement team monitoring the database. Systems which effectively extract information from unstructured free-text documents using natural language processing (NLP), rather than requiring clinicians to overstructure data at the time of input, can augment existing workflows and provide the backbone for these centralized results management tools.

In this manuscript, we develop and describe a comprehensive results management tool as applied to a particular incidental finding—adrenal incidentalomas. An adrenal incidentaloma is defined as a mass greater than 1 cm in diameter discovered on an image obtained for reasons other than to evaluate the adrenal glands.[1] Adrenal incidentalomas are both common and potentially pathologic, making them a valuable target for improved management.[2,12] Approximately 4.4% of all abdominal computed tomography (CT) scans contain incidental adrenal gland lesions[12]; this percentage rises to nearly 10% for elderly patients.[2] As the use of imaging increases[9] and the population ages,[13] the number of masses that need tracking and follow-up care will increase substantially. The majority of these lesions are found to be benign and nonfunctional, but some types of masses require intervention, including primary adrenal cancers (e.g., adrenocortical carcinoma), metastatic lesions, hormone secreting tumors, and life threatening pheochromocytomas.[1,2,14] The true percentage of pathologic adrenal incidentalomas is not well established, but is estimated to be 20% when accounting for both pathologic biochemical dysfunction and malignancy.[15]

The consensus from several specialty organizations is that all patients with adrenal incidentalomas be followed up for the possibility of malignancy and subclinical hormonal hyperfunction,[2,16–19] at least with the patient's PCP and potentially with an endocrinologist or endocrine surgeon. Unfortunately, the majority of patients do not receive appropriate follow-up. Among incidental adrenal lesions, follow-up is especially low, likely because it involves a combination of radiographic studies and multiple biochemical evaluations to rule out pheochromocytoma and autonomous secretion of aldosterone or cortisol. One recent study found that incidental adrenal masses are appropriately followed up less than 10% of the time.[20] An additional study found that only 18.4% of patients with adrenal incidentalomas received a complete initial evaluation of the mass per American Association of Clinical Endocrinologists and American Association of Endocrine Surgeons guidelines.[21] Adrenal incidentalomas exemplify many of the problems with long-term incidental results management identified above, and therefore represent an appropriate first target for an improved results management system.

Objectives

In this manuscript, we employ NLP to identify radiology reports which contain mentions of *previously unseen* adrenal incidentalomas. On top of this information extraction backbone, we develop a back-end database and front-end web application, designed to improve longitudinal coordination of incidentaloma management. The application, to be deployed at a large metropolitan safety net hospital, maintains and displays a list of incidentalomas requiring follow-up, along with a single-page dashboard of relevant clinical data (e.g., laboratories, appointments), and enables communication and clinical management of these findings.

Methods

Dataset

Our dataset consisted of 4,090 radiology reports drawn from our institution's database of reports from 2016 to 2018. Of these, 251 reports already known to contain adrenal incidentalomas generated from existing clinical workflows from January 1 to December 31, 2016 were included and a random sample of 3,839 additional reports were hand labeled. The existing clinical workflow used a simple string matching protocol (matching any report containing "adrenal lesion," "adrenal nodule," "adrenal mass," or "adrenal adenoma") followed by manual screening of the returned results. We included the sample of already known positive reports in addition to our random sample to ensure a sufficient number

of positive examples to effectively train the model. This addition of known positive reports functionally acts as a form of dataset augmentation by oversampling positive examples of new adrenal incidentalomas, improving the model's ability to learn semantically useful features of positive reports. As the rate of new adrenal incidentalomas among all new radiology reports is likely smaller than the rate in our dataset, this form of data augmentation increases the likelihood of identifying infrequent positive reports. The reports consisted of a broad set of imaging modalities—CT, magnetic resonance imaging (MRI), and positron emission tomography-computed tomography (PET-CT)—of the chest, abdomen, and pelvis. We did not limit our dataset to abdominal imaging studies because adrenal findings were often observed on chest imaging. Radiology reports for scans ordered specifically to evaluate the adrenal glands were excluded. The report corpus includes a variety of reporting templates and patients of all ages and genders. The study was approved by our institution's Institutional Review Board.

Annotation

Our data were annotated at the document level by a PGY-4 surgery resident and post-doctoral research fellow. Annotations were reviewed with a fellowship-trained adrenal surgeon and a fellowship-trained body imaging radiologist was consulted to answer questions as needed. A report was assigned a "true" label if it (1) contained textual evidence of at least one incidental adrenal finding which had not been observed before, (2) was greater than 1 cm, and (3) was discovered on an institutional imaging study and not at an outside hospital. A report with an incidental finding that had been noted previously was not assigned a "true" label. For instance, a report with a single incidental finding described as, e.g., "stable," "unchanged," or "previously observed" was given a "false" label. Each report was labeled by one of the two authors (T.F. and J.S.).

Neural Network Design and Training Protocol

The structure of the task is such that the majority of text in a given report is not relevant to the document output label (i.e., most of the report is not concerned with the adrenal glands) mirroring feature detection tasks found in computer vision. For the purposes of document classification in this context, the region of interest that determines whether a new adrenal incidentaloma is identified is roughly represented by certain groups of words clustered together. We therefore opted to use a simple shallow convolutional neural network (CNN), which is classically used for feature detection tasks and has been used for text classification tasks in the past,[22] including a 2019 study on identifying incidental findings in radiology reports.[23] Our decision to use a convolutional, rather than recurrent neural network (RNN), for text classification also relates to the efficiency of these networks for the stated feature detection task. Less computationally intensive than RNNs, CNNs can run effectively with limited resources. The improvement in model training time and inference on a machine without graphics processing units (GPUs) is not universal,[24] but we empirically determined that this principle held true for the purpose of identifying new adrenal incidentalomas. This makes model implementation feasible in a resource constrained hospital setting, the final goal of the work presented here.

Our network used pretrained 300-dimensional Global Vector (GloVe) word embeddings derived from the Common Crawl web dataset[25] and packaged within the SpaCy Python package. Documents were tokenized using the built-in SpaCy tokenizer.[26] Although more task-specific word embeddings exist,[23] our network was trained with GloVe word embeddings to improve model robustness and ensure similar model performance across institutions. ►Appendix A.1 contains further discussion on the decision to use GloVe word embeddings. The network itself consisted of a single convolutional layer for feature detection (200 units, convolutional kernel size 7) followed by a global max-pooling layer, which selects the highest feature activation for each convolutional filter from across the entire document length. A two layer, fully connected neural network, with a ReLu nonlinearity layer in between the two dense layers, was applied to the output of the max pooling layer and a softmax was applied to produce the final binary output. A dropout rate of 0.3 was applied after the convolutional and first dense layers. The hyperparameter selection strategy is described in further detail in ►Appendix A.2.

Labeled documents were split into training, validation, and test sets (80%/10%/10%). Training was performed using the Adam algorithm[27] with a learning rate of $1e-4$. Training was continued until validation loss ceased to decrease with a patience of three epochs. Final performance was calculated on the test set. Model performance is reported in conjunction with confidence intervals, sensitivity, and specificity. The method for the calculation of these measures is elaborated in ►Appendix A.3.

Model Interpretability

Clinical machine learning systems must be interpretable by human users.[28] Many approaches for improving interpretability among "black box" systems have been developed in the machine learning literature; for this study, we opted to use the Gradient Class Activation Maps (Grad-CAM) algorithm, which identifies parts of an input which were important in the model's final decision about that input.[29] We used Grad-CAM to generate saliency maps for our model's predictions to visualize the importance of each word token in the model's decision making process. By design, the Grad-CAM layer does not affect model output or performance.

Web Application

To implement our system into the clinical workflow, we built a custom web application that uses the trained neural network model to generate predictions and provides a dashboard to enable risk stratification and prioritization, tracking, and clinical management of identified findings (e.g., to track follow-up appointment scheduling, PCP notification).

To accomplish this, we maintain a database of reports on a secure internal web server. Our system receives periodic transfers of data consisting of all radiology reports for recent imaging studies and uses a scheduled task to process the new

Table 1 Description of dataset

Total reports	**4,090**
New adrenal incidentalomas	
Reports with new adrenal incidental findings	404
Reports without new adrenal incidental findings	3,686
Imaging study type (among reports with new incidentalomas)	
CT abdomen/pelvis (with or without contrast)	225
CT chest/abdomen/pelvis	37
CT chest	104
CT abdomen	15
CT chest/abdomen	2
MRI abdomen	10
Other	11

Abbreviations: CT, computed tomography; MRI, magnetic resonance imaging.

batch of reports with the neural network model. The model generates a new set of machine labels, which are stored in the database. Simple rules are used to identify relevant laboratories (e.g., ACTH, aldosterone, etc.), clinical appointments (e.g., with endocrinology or surgery), and orders (e.g., radiologic studies), extract them from our institution's EMR database, and display all of the information on a single dashboard. This minimizes the clicks required to view and synthesize the relevant information.

Results

Dataset

Our dataset included 404 positively labeled reports (9.9%) and 3,686 (90.1%) negatively labeled reports (►Table 1). Among positively labeled reports, 166 (55.5%) came from reports on CT studies of the abdomen/pelvis, while 73 (24.4%) were discovered on CT scans of the chest. Of note, many of these reports contained either negation phrases (e.g., "no adrenal nodules") or phrases indicating that the incidental finding had been previously observed; both of these scenarios were labeled as negative.

Model Performance

On the unseen test set, our network achieved an accuracy of 97.3% (95% confidence interval [CI]: 95.2–98.5%), a recall (sensitivity) of 92.9% (80.9–97.5)%, a precision (positive predictive value) of 83.0% (69.9–91.1)%, a specificity of 97.8% (95.8–98.9)%, and an F1 score (harmonic mean of recall and precision) of 87.6%. Hyperparameter selection strategy is discussed in further detail in ►Appendix A.2**.** Model performance on the validation set, as well as 2 × 2 frequency tables outlining model validation and test set performance in comparison to ground truth, expert human labels, are available in ►Appendices B and C. The model was trained to convergence in less than 30 minutes on a standard institutional desktop with no specialized computing hardware (i.e., no graphics processing units [GPUs]), demonstrating the feasibility of such an approach in a non-GPU-resourced environment. On this machine, the system is capable of processing thousands of documents per minute.

Table 2 Text spans with the highest Grad-CAM scores for a randomly selected set of reports

Report	Grad-CAM spans
1	"3. Left adrenal nodule is incompletely" "Is a left adrenal nodule. Bones,"
2	"mm hypodense left adrenal nodule that is"
3	"of the left adrenal, there is"
4	"mass. Recommend adrenal MRI for further"
5	"1.3-cm left adrenal adenoma. There"
6	"INFORMATION: adrenal mass noted on"
7	"1.7-cm right adrenal nodule which measures," "likely represents an adrenal adenoma. The"
8	"there are incompletely characterized 1.4 cm"
9	"1.6 cm-right adrenal nodule which measures"
10	"consistent with an adrenal adenoma. Addendum"

Abbreviations: Grad-CAM, gradient class activation maps; MRI, magnetic resonance imaging.

Interpretability

We ran the Grad-CAM algorithm over all of the reports predicted to contain a new adrenal incidentaloma. This procedure generated a set of text spans which were most "important" to the model's predictions, based on the gradients of the final prediction with respect to the feature maps of the convolutional layer. ►Table 2 shows the text spans with the highest Grad-CAM scores for a randomly selected set of positively predicted reports.

Web Application

The web application displays a list of patients with newly identified findings and includes the most relevant text from each report as identified by the Grad-CAM algorithm. The application is intended for the hospital's quality department, but will also be accessible to individual clinicians. It allows for a variety of user actions, including: (1) collaborative note taking for a particular finding (e.g., follow-up appointment scheduling, sending a letter to a patient), (2) marking a finding as closed (i.e., when adequate follow-up has been performed), and (3) confirmation or correction of machine predictions. When a user verifies a machine prediction as correct, or marks it as incorrect, this information is stored as a "human label," which supersedes any machine predicted labels. This not only allows for correction of machine mistakes within the system, but also increases the size of the labeled training set. These labels are combined with the initial label set; as these new labels accumulate, the model will be retrained to further improve performance. ►Fig. 1 provides an overview of the complete application development process. A screenshot of the web application is shown in ►Fig. 2.

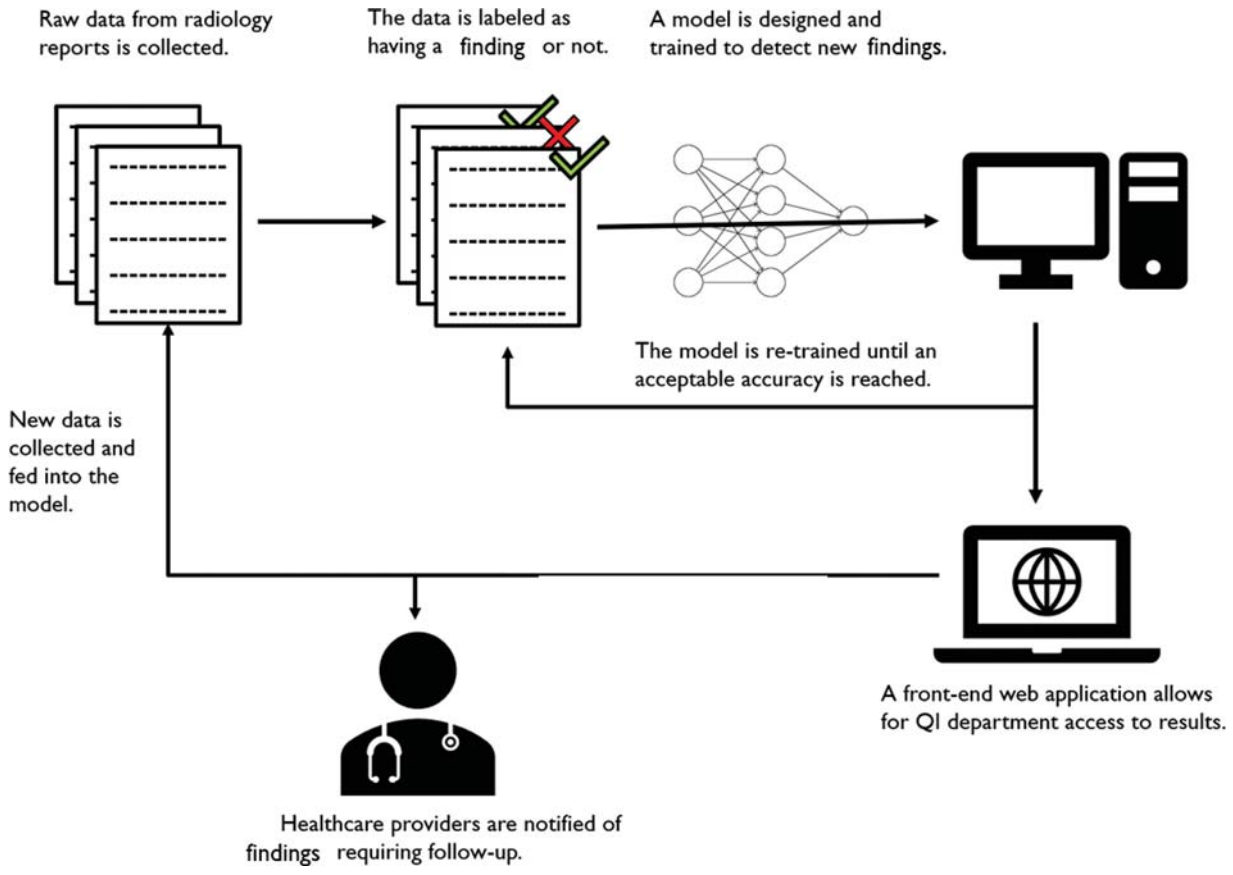

Fig. 1 Overview of the framework required to develop a machine-learning enabled web application to track incidental findings. Once all the documents are collected, the reports are labeled and are used to train a machine learning classification algorithm. The machine learning algorithm can prospectively evaluate new reports and the results are available as a front-end web application to facilitate the process of alerting appropriate providers of the incidental finding.

Discussion

In this study we trained a model to identify radiology reports containing notation of newly discovered adrenal incidentalomas and developed a software tool that leverages this model to enable improved management of these findings. Our model achieved a recall (sensitivity) of 100%, a precision of 84.8%, and an F1 of 91.8% on the task of identifying previously unseen radiology reports that contain textual evidence of a newly discovered adrenal incidentaloma. Furthermore, the inclusion of a user verification active learning component in the application enables continued model refinement.[30] This system performed comparably to expert level human performance, and importantly, has significant resource saving potential both in terms of time and money.

Clinical Significance

Currently, the standard protocol for long-term care coordination of adrenal incidentalomas is provider-to-provider communication, which is highly error prone. To our knowledge, the automated system presented here is the first created for tracking this information prospectively.

Consider a protocol that does not utilize automated tools: (1) The reading radiologist documents the incidentaloma in the radiology report, (2) the ordering physician reads the report and may (or may not) communicate the incidental findings to the patient and PCP, who then takes appropriate follow-up action—shared decision making with the patient, interval imaging, further laboratory testing, or referral to a specialist who performs the evaluation. A second such pathway exists for more centralized management of incidental findings: (1) The reading radiologist documents the incidentaloma in the radiology report; (2) A central coordinator reads all radiology reports, or a subset generated via a keyword search, identifies the incidentaloma, and then (3) schedules appropriate follow-up with patient's PCP or specialist who then takes appropriate follow-up action. In either case, these protocols would require a significant amount of human effort and lead to many transitions of care, propagating errors inherent in such handoffs. ►Fig. 3 displays these pathways, as well as the more direct pathway made possible through automated methods, such as the one presented here using NLP embedded within a custom web application.

We designed the web application around the NLP model's output to allow streamlined management of these important clinical findings at our institution without requiring physicians to input additional structured data. This framework is simple, readily extendible, and a meaningful improvement over current practice. An overview of the steps required to train and deploy a machine learning model to identify

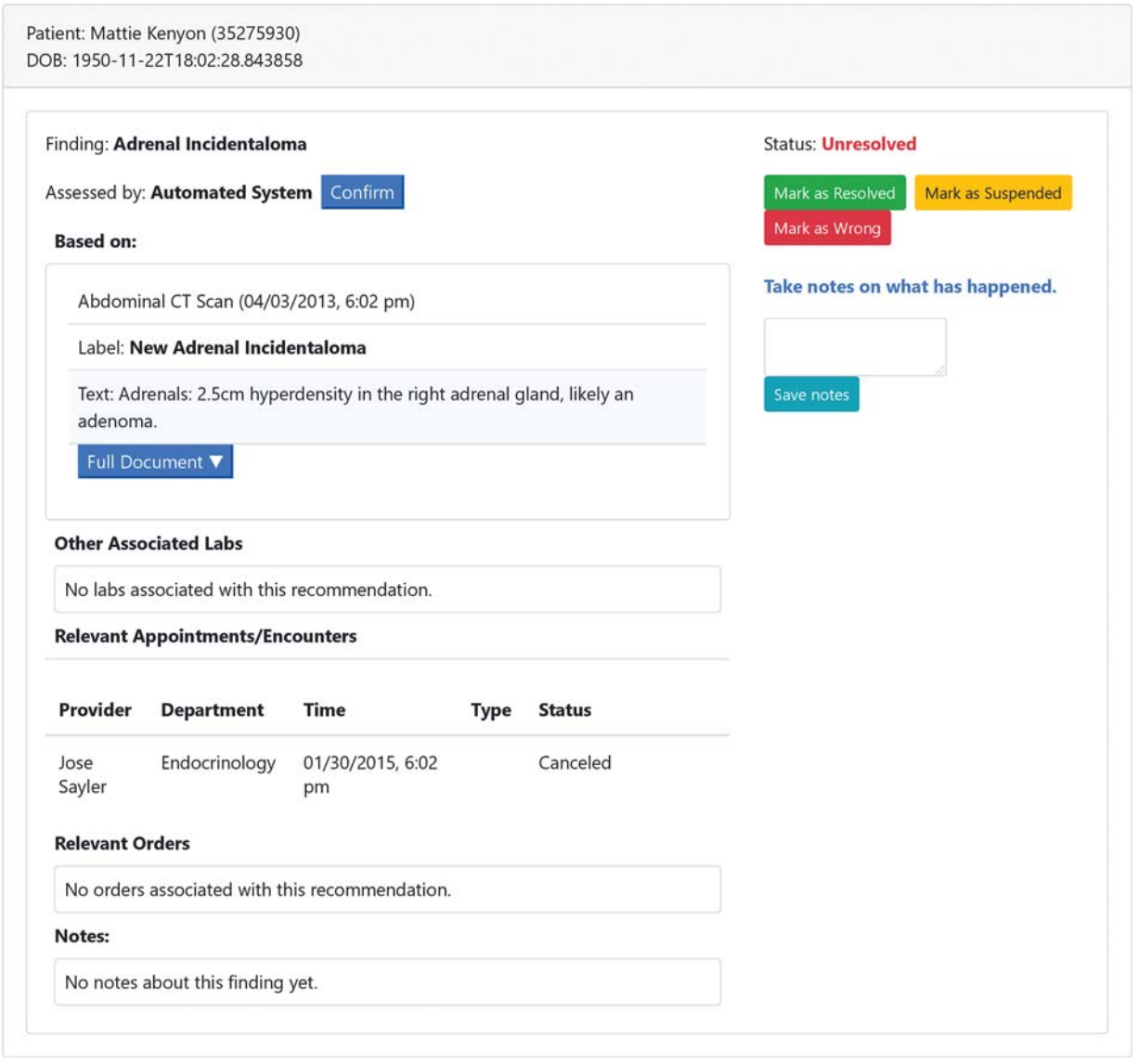

Fig. 2 Screenshot of results management web application running on simulated data (all data are fictional, not real). The application displays the list of adrenal incidentalomas detected by the machine learning algorithm. The displayed text is the full sentence surrounding the Grad-CAM predicted tokens that identify this document as one containing a new adrenal incidentaloma; the full document can also easily be viewed. The application allows multiple concurrent users to view relevant clinical data (e.g., laboratories, upcoming appointments, orders), manage the list over time, and make any necessary corrections to the machine predictions. Grad-CAM, gradient class activation maps.

incidental findings using this framework is shown in ►Fig. 1. The iterative process of tool development, deployment, and improvement will take place in real time with physicians and quality department staff. This tool is meant to augment the institution's ability to track and manage these findings without workflow interruptions for frontline physicians, who are already time constrained in patient interactions and chart review. Efforts to standardize the clinical judgements of a diverse group of physicians who have had different training and habits may not be fast enough to deliver care improvement at a sufficient pace.[31] Though we will make the tool available to interested PCPs, the target users work in the quality department, which is particularly well suited for initial management of these kinds of high-risk findings. As demonstrated here, the automated identification of incidental findings, such as adrenal incidentalomas, represents an opportunity to improve patient care and outcomes without negatively impacting the workflow of providers.

Prior Approaches

Attempts to automatically identify imaging findings in radiology reports requiring follow-up date back as far as 1993.[32] These prior efforts have explored the utility of an automated identification tool across various modalities, including chest X-ray, CT, MRI, and ultrasound.[33–44] These studies used rule-based systems, non-neural machine learning methods, or both, to identify textual evidence of findings that require follow-up either across all organ systems or focused on a particular organ. A 2019 study successfully used a CNN to identify incidental findings in radiology reports, supporting the idea that interactive NLP applications can be adopted as part of routine clinical care.[45]

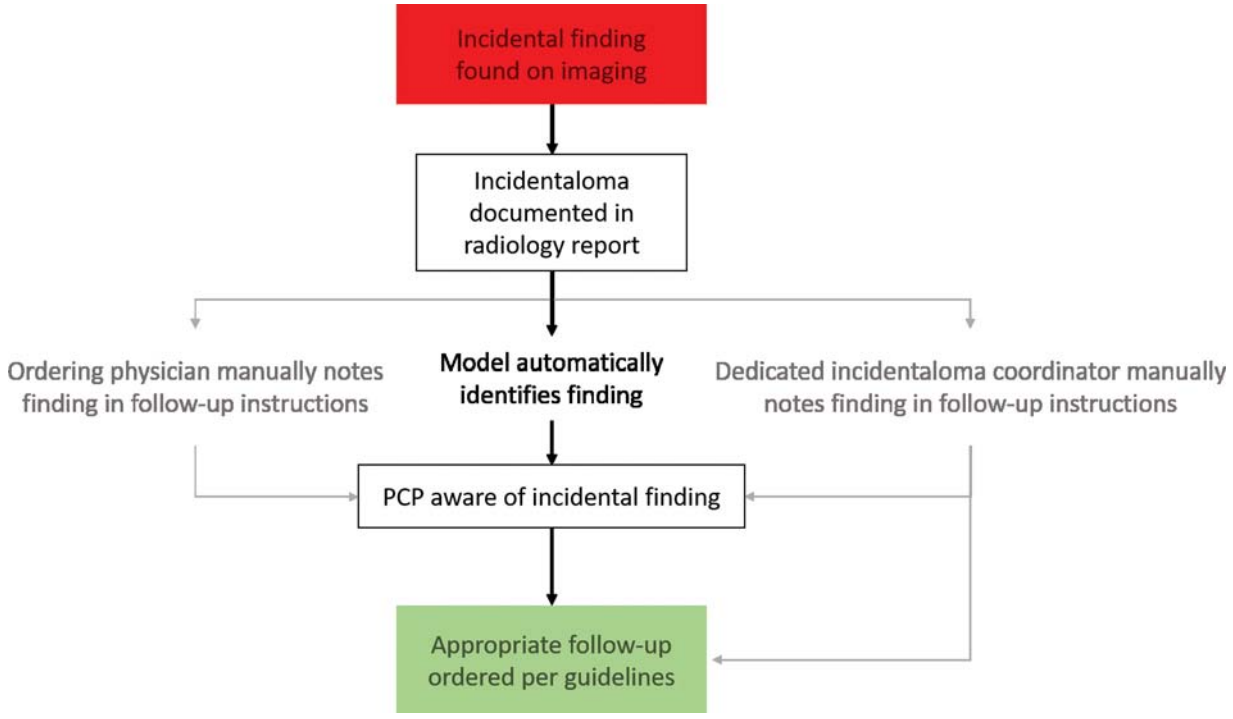

Fig. 3 Options for ensuring proper follow-up of incidentaloma. Incidentalomas recorded in the radiology report can be communicated to appropriate providers by direct action of the ordering physician, employment of an incidentaloma coordinator specifically examining radiology reports for evidence of new incidentalomas, or with the use of an automated identification tool such as the one described in our study.

Traditional machine learning approaches require the manual development of highly tuned imaging features. When the dataset used to train these models is expanded to include more institutions, note types, or EMRs, these features must undergo nontrivial recalibration and model performance may suffer significantly. Preliminary analyses of rule-based approaches on our classification task produced significantly worse results, largely due to negation and temporal statements (e.g., comments such as "as previously measured" or "stable," which we aimed to exclude). Unlike these prior efforts, we opted for a neural network solution, which easily incorporates new unstructured data to improve performance, thereby increasing model generalizability and avoiding bottlenecks in model development and deployment. Rather than attempt to produce an exhaustive set of rules, we produced additional document annotations to increase the size of our dataset; these annotations are not time consuming to produce and are sufficient to train a neural network to perform accurately. Therefore, the neural network core of our software makes extension to other institutions simple and likely to provide performance gains at our institution or others.

Interpretability

We can use our model's Grad-CAM outputs to qualitatively assess its predictions. As ►Table 2 demonstrates, the model is attending to sentences relating to the adrenal nodules, suggesting that the system has learned to use the "correct" parts of the document (i.e., those relating to statements about adrenal incidentalomas) for classification. These examples offer additional evidence that the model has been optimized to identify higher order semantic features of the text that correspond to new adrenal incidentalomas, rather than simply relying on string matching or other easily exploitable clues attributable to the dataset itself. In addition to allowing for verification of machine predictions, the Grad-CAM outputs may prove helpful to improve the clinician user experience in the software, by allowing them to quickly focus on the parts of the document which contain additional contextual information (e.g., size, location) about the identified adrenal incidentaloma.

Limitations

There are several limitations to our study. First, there are a couple limitations pertaining to dataset curation. Although a fellowship-trained radiologist was consulted when report language was unclear, positive imaging findings noted in a report were not further verified by a radiologist. Multiple physicians were involved in the annotation procedure to produce expert level annotations, but because each report was documented by a single annotator, the calculation of interrater reliability metrics (e.g., Cohen's kappa) was not possible. Second, while our study achieves an impressive 91.8% F1 measure on the test set, our dataset is limited to a single academic institution. We employed data augmentation techniques to improve the model's ability to recognize positive examples, but the rate of positive examples in real hospital data is likely smaller. The sensitivity of the model should not be impacted in the setting of reduced disease prevalence, but the PPV may be lower. Other institutions of varying types may introduce additional complexities that limit our tool's usefulness in a wider setting, though this risk

is mitigated by the underlying structure of our model, which favors generalizability. Third, the identification of new adrenal incidentalomas in radiology reports represents only one of many such kinds of incidentally discovered findings requiring clinical follow-up. Given the globally poor adherence to recommendations regarding incidental findings in radiology reports, further work is needed to address other findings that would benefit from automated identification. Lastly, as our purpose was to develop a lightweight, accurate, and easily generalizable machine learning model to address gaps in clinical continuity, we have not compared our choice of model, CNN, to other models which could result in performance gains.

Future Directions

In future work we intend to expand the dataset to include reports from other institutions to build a more readily generalizable model. We also plan to expand the scope of the software and underlying NLP models to include additional clinically actionable radiological findings (e.g., solitary pulmonary nodules). We will iterate on the front-end interface in collaboration with users and improve integration with existing EMR systems—ideally these actionable clinical items would be pushed back into the institution's EMR. Further descriptive characteristics of the incidentalomas can also be identified and extracted from these radiology reports. Though our model is able to identify new adrenal incidentaloma findings, it does not extract specific descriptors of the nodule (e.g., size or appearance), nor the follow-up measures specified in the report. Extracting these descriptors can facilitate the development of an automated pipeline to automatically schedule appropriate follow-up, reducing the burden on providers as well as opportunities for human error. Lastly, we intend to prospectively address the impact of this tool on rates of clinical follow-up following a period of time during which it is in active use.

Conclusion

In conclusion, we demonstrate the feasibility of training a neural network to automatically identify free-text radiology reports that contain textual evidence of previously unseen adrenal incidentalomas. We achieve accurate, clinically useful performance using a small training corpus, and train the network in less than 1 hour on a machine with no GPU. We also develop a front-end web application for centralized tracking and management of incidentalomas by the hospital's quality department. The workflow enabled by our system does not require changing the practice patterns of radiologists, which can impede efforts at automating the detection, reporting, and communication of incidental radiographic findings. Furthermore, the overall system is designed to be generalizable to any similar task (e.g., solitary pulmonary nodules, mammographic findings, etc.). Although neural network models have been employed in medicine, relatively few actionable tools have been created to successfully harness their predictive power in real time. With the creation of our neural network enabled web application, we demonstrate that these innovations can be put to use today to fill clinical care gaps that would otherwise be burdensome to implement.

Clinical Relevance Statement

Machine learning models using neural networks can be deployed in clinical practice to automate otherwise onerous chart review tasks without specialized computing hardware. High-risk incidental findings such as adrenal incidentalomas can be tracked and managed with a centralized web application. Practitioners and health care systems should consider deployment of such tools for quality improvement and risk reduction purposes.

Multiple Choice Questions

1. How will automated identification of incidental findings affect the individual workflow of the ordering provider?

a. The provider will communicate findings directly to a dedicated incidental finding coordinator.
b. The provider will communicate findings directly to the patient's primary care provider.
c. The provider will confirm suspected incidental findings in radiology reports.
d. No adjustment of individual workflow is required.

Correct Answer: The correct answer is option d. One of the greatest benefits of automated identification tool is that they do not require additional input from providers. Our application is able to monitor radiology reports in the background, and new pipelines can be developed that leverage the model's prediction to order appropriate follow-up without direct intervention (d). Without automated identification, communication of incidental findings requires active patient handoff measures, such as the (a) the employment of a dedicated coordinator who searches through radiology reports for evidence of new incidental findings or (b and c) physicians closely reading and relaying of information to a patient's primary care provider.

2. Which of the following statements regarding the epidemiology and impact of adrenal incidentalomas is true?

a. The majority of adrenal incidentalomas are indicators of serious underlying pathology.
b. As many as 1 in 10 abdominal CT scans contain incidental adrenal gland lesions in certain patient populations.
c. Existing estimates of appropriate follow-up for adrenal incidentalomas generally exceed 80%.
d. The projected number of reported incidental adrenal findings is expected to remain stable over time.

Correct Answer: The correct answer is option b. The overall prevalence of adrenal incidentalomas is 4%, but this rate rises to nearly 10% in elderly populations (b). Although the majority of incidental adrenal findings are benign and nonfunctional (a), they may be pathological by means of an underlying malignancy or functional hormonal secretion. Given that the number of reported incidental adrenal findings is expected to rise significantly as a result of an increased

reliance on imaging results for treatment decisions and an aging population (c), it is important that we optimize methods for ensuring appropriate follow-up. Current rates of follow-up measures of incidental findings are suboptimal, hovering around 50% (d).

3. Which of the following is true regarding the machine learning model employed in this study?

a. The features used by the model to make predictions were selected based on finely tuned parameters following manual review of the dataset.
b. The model requires the use of a graphics processing unit (GPU) to make predictions within a reasonable timeframe.
c. The model uses a shallow neural network to automatically identify important text spans in radiology reports that support its predictions.
d. There are currently no methods available for providing qualitative insight into how the model makes decisions.

Correct Answer: The correct answer is option c. Unlike traditional machine learning models or rule-based approaches to automated prediction, neural network models such as the one in this study (c) can learn to automatically identify and use the most salient features in the data for the prescribed task. Earlier machine learning approaches require the careful derivation of useful text features (a) that must then be adapted as the dataset complexity increases. Our model intentionally uses a shallow CNN architecture to ensure that calculations can be completed in real time without the need for a GPU (b). To gain further insight into model function, we use gradient class activation maps, which highlight the text spans the model pays most attention to when making predictions (d).

Protection of Human and Animal Subjects

This study was reviewed and approved by the Boston Medical Center and Boston University Medical Campus Institutional Review Board (IRB).

Conflict of Interest

This project was initiated by the Department of Quality and Patient Safety at Boston Medical Center; however, four members of the research team (W.B., J.S., J.K., and A. S.), who participated in this work as medical students, are also part-owners of River Records, LLC, a health care informatics company with the goal of developing tools that analyze clinical free-text data. There are no existing financial relationships between River Records and Boston Medical Center, nor does the company currently have any contracts, employees, or products. None of the coding or natural language processing techniques described in this paper represent patentable intellectual property and readers are free to reproduce and expand upon this work.

Acknowledgments

The authors wish to thank Stephanie Talutis, Praveen Sridhar, Megan Janeway, Chelsea Vigna, Katherine Valles, and Melissa Griswold for their efforts in supporting the success of this study.

References

1 Young WF Jr. Clinical practice. The incidentally discovered adrenal mass. N Engl J Med 2007;356(06):601–610
2 Terzolo M, Stigliano A, Chiodini I, et al; Italian Association of Clinical Endocrinologists. AME position statement on adrenal incidentaloma. Eur J Endocrinol 2011;164(06):851–870
3 Young WF Jr. Management approaches to adrenal incidentalomas. A view from Rochester, Minnesota. Endocrinol Metab Clin North Am 2000;29(01):159–185
4 Mantero F, Terzolo M, Arnaldi G, et al; Study Group on Adrenal Tumors of the Italian Society of Endocrinology. A survey on adrenal incidentaloma in Italy. J Clin Endocrinol Metab 2000;85 (02):637–644
5 Cawood TJ, Hunt PJ, O'Shea D, Cole D, Soule S. Recommended evaluation of adrenal incidentalomas is costly, has high false-positive rates and confers a risk of fatal cancer that is similar to the risk of the adrenal lesion becoming malignant; time for a rethink? Eur J Endocrinol 2009;161(04):513–527
6 Mabotuwana T, Hombal V, Dalal S, Hall CS, Gunn M. Determining adherence to follow-up imaging recommendations. J Am Coll Radiol 2018;15(3 Pt A)422–428
7 Dang PA, Kalra MK, Blake MA, et al. Natural language processing using online analytic processing for assessing recommendations in radiology reports. J Am Coll Radiol 2008;5(03):197–204
8 Sistrom CL, Dreyer KJ, Dang PP, et al. Recommendations for additional imaging in radiology reports: multifactorial analysis of 5.9 million examinations. Radiology 2009;253(02):453–461
9 Bhargavan M, Sunshine JH. Utilization of radiology services in the United States: levels and trends in modalities, regions, and populations. Radiology 2005;234(03):824–832
10 Weingart SN, Yaghi O, Barnhart L, et al. Preventing diagnostic errors in ambulatory care: an electronic notification tool for incomplete radiology tests. Appl Clin Inform 2020;11(02):276–285
11 O'Connor SD, Khorasani R, Pochebit SM, Lacson R, Andriole KP, Dalal AK. Semiautomated system for nonurgent, clinically significant pathology results. Appl Clin Inform 2018;9(02):411–421
12 Bovio S, Cataldi A, Reimondo G, et al. Prevalence of adrenal incidentaloma in a contemporary computerized tomography series. J Endocrinol Invest 2006;29(04):298–302
13 Ortman JM, Velkoff VA, Hogan H. An Aging Nation: The Older Population in the United States. Suitland-Silver Hill, MD: United States Census Bureau; 2014
14 Grumbach MM, Biller BM, Braunstein GD, et al. Management of the clinically inapparent adrenal mass ("incidentaloma"). Ann Intern Med 2003;138(05):424–429
15 Nieman LK. Approach to the patient with an adrenal incidentaloma. J Clin Endocrinol Metab 2010;95(09):4106–4113
16 Zeiger MA, Thompson GB, Duh Q-Y, et al. American Association of Clinical Endocrinologists; American Association of Endocrine Surgeons. The American Association of Clinical Endocrinologists and American Association of Endocrine Surgeons medical guidelines for the management of adrenal incidentalomas. Endocr Pract 2009;15(Suppl 1):1–20
17 Fassnacht M, Arlt W, Bancos I, et al. Management of adrenal incidentalomas: European Society of Endocrinology Clinical Practice Guideline in collaboration with the European Network for the Study of Adrenal Tumors. Eur J Endocrinol 2016;175(02): G1–G34
18 Kapoor A, Morris T, Rebello R. Guidelines for the management of the incidentally discovered adrenal mass. Can Urol Assoc J 2011;5 (04):241–247
19 Lee JM, Kim MK, Ko SH, et al. Korean Endocrine Society, Committee for Clinical Practice Guidelines. Clinical guidelines for the management of adrenal incidentaloma. Endocrinol Metab (Seoul) 2017;32(02):200–218
20 Feeney T, Talutis S, Janeway M, et al. Evaluation of incidental adrenal masses at a tertiary referral and trauma center. Surgery 2020;167(05):868–875

21 Becker J, Woloszyn J, Bold R, Campbell MJ. The adrenal incidentaloma: an opportunity to improve patient care. J Gen Intern Med 2018;33(03):256–257

22 Kim Y. Convolutional neural networks for sentence classification. Paper presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: Association for Computational Linguistics; 2014:1746–1751

23 Trivedi G, Hong C, Dadashzadeh ER, Handzel RM, Hochheiser H, Visweswaran S. Identifying incidental findings from radiology reports of trauma patients: an evaluation of automated feature representation methods. Int J Med Inform 2019;129:81–87

24 Shi S, Wang Q, Xu P, Chu X. Benchmarking state-of-the-art deep learning software tools. Paper presented at: 2016 7th International Conference on Cloud Computing and Big Data (CCBD). Macau: IEEE; 2016

25 Pennington J, Socher R, Manning CGlove: global vectors for word representation. Paper presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP);. 2014. Doha: Association for Computational Linguistics

26 spaCy. Industrial-strength natural language processing in python.. Available at: https://spacy.io/. Accessed September 7, 2019

27 Kingma DP, Ba JAdam: a method for stochastic optimization. arXiv [csLG]. December 2014. Available at: http://arxiv.org/abs/1412.6980

28 Ahmad MA, Teredesai A, Eckert CInterpretable machine learning in healthcare. Paper presented at: 2018 IEEE International Conference on Healthcare Informatics (ICHI);. 2018. Hong Kong: IEEE Computer Society:447–447

29 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra DGrad-CAM: visual explanations from deep networks via gradient-based localization. arXiv [csCV]; October 2016. Available at: http://arxiv.org/abs/1610.02391. Accessed June 23, 2020

30 Hanneke S. Rates of convergence in active learning. Ann Stat 2011; 39(01):333–361

31 Lim PS, Schneider D, Sternlieb J, et al. Process improvement for follow-up radiology report recommendations of lung nodules. BMJ Open Qual 2019;8(02):e000370

32 Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016;279 (02):329–343

33 Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Comput Biomed Res 1993;26(05):467–481

34 Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. J Biomed Inform 2013;46(05):869–875

35 Dutta S, Long WJ, Brown DFM, Reisner AT. Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. Ann Emerg Med 2013;62(02):162–169

36 Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. Automatic identification of critical follow-up recommendation sentences in radiology reports. AMIA Annu Symp Proc 2011;2011:1593–1602

37 Meng X, Ganoe CH, Sieberg RT, Cheung YY, Hassanpour S. Assisting radiologists with reporting urgent findings to referring physicians: a machine learning approach to identify cases for prompt communication. J Biomed Inform 2019;93:103169

38 Meng X, Heinz MV, Ganoe CH, Sieberg RT, Cheung YY, Hassanpour S. Understanding urgency in radiology reporting: identifying associations between clinical findings in radiology reports and their prompt communication to referring physicians. Stud Health Technol Inform 2019;264:1546–1547

39 Pham A-D, Névéol A, Lavergne T, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics 2014;15:266

40 Hassanpour S, Bay G, Langlotz CP. Characterization of change and significance for clinical findings in radiology reports through natural language processing. J Digit Imaging 2017;30(03): 314–322

41 Mabotuwana T, Hall CS, Dalal S, Tieder J, Gunn ML. Extracting follow-up recommendations and associated anatomy from radiology reports. Stud Health Technol Inform 2017;245:1090–1094

42 Dalal S, Hombal V, Weng W-H, et al. Determining follow-up imaging study using radiology reports. J Digit Imaging 2020;33 (01):121–130

43 Grundmeier RW, Masino AJ, Casper TC, et al; Pediatric Emergency Care Applied Research Network. Identification of long bone fractures in radiology reports using natural language processing to support healthcare quality improvement. Appl Clin Inform 2016;7(04):1051–1068

44 Sevenster M, Buurman J, Liu P, Peters JF, Chang PJ. Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports. Appl Clin Inform 2015;6(03):600–110

45 Trivedi G, Dadashzadeh ER, Handzel RM, Chapman WW, Visweswaran S, Hochheiser H. Interactive NLP in clinical care: identifying incidental findings in radiology reports. Appl Clin Inform 2019;10 (04):655–669

Appendix A

1. Word Embedding Selection

In our study, we prioritized the development of immediately deployable, actionable, clinically-useful applications in a setting with limited computational resources. For these reasons, we decided to use a well-known, well-studied, publicly-accessible set of vector embeddings. Global Vector (GloVe) pretrained vectors have been found to perform well at a wide range of NLP tasks. Although other pretrained vector sets may be more specific to the task of analyzing radiology reports (such as those developed by a 2019 study[23]), they are not validated in multiple settings over time as GloVe vectors have been. The use of a well-known, widely-used set of pretrained vectors, such as GloVe vectors, allowed us to increase the robustness of our model. In addition, because GloVe vectors are trained on significantly larger corpora, they are more robust to generalization tasks. In this manuscript, we demonstrate one example of a machine-learning powered application at a large institution. We chose GloVe vectors to allow for simpler emulation by investigators and quality improvement leaders at other institutions; its use should produce similar performance across different institutions as compared with word vectors that were trained on less data but dedicated to more specific tasks.

2. Hyperparameter Selection Strategy

The number of layers were selected based on the notion that the features of import, intuitively, were unlikely to span more than seven words and thus did not require multiple convolutional layers to agglomerate information beyond this length. The dropout rate was selected based on similar studies, such as a 2014 study using convolutional neural networks (CNNs) for sentence classification.[22] We opted not to do an exhaustive hyperparameter search, as our goal was to find a simple (i.e., not computationally expensive) model which performed sufficiently well for our purposes, i.e., to provide a substantial improvement over the current system, which consists of no standardized process for free-text clinical results tracking. Furthermore, more complex models (including deeper CNNs as well as long short term memory) generally require more labeled data to train effectively, suggesting that it made sense to start with a simpler model and add complexity as necessary, rather than starting with a more complex model. Increasing the size of the labeled dataset proved to be the most effective intervention to improve performance, and based on our past experience it is likely that adding more labeled examples to learn from would be more effective than searching for optimal model hyperparameters.

3. Calculation of Confidence Intervals, Sensitivity, and Specificity

Confidence intervals were derived based on a single trained model evaluated on the test set once (with a single 80/10/10 split), and are meant to reflect a confidence interval for this model's performance on unseen data drawn from the same distribution as the test set. They are not meant to be interpreted as confidence intervals for the general training process or model architecture, the way that *k*-fold cross validation would be.

The decision to use a single holdout test set, as opposed to model evaluation with cross validation is twofold. First, the use of repeated cross validation to assess the performance of a model in a particular practical domain may produce misleading results. The test set, which acts as the set of strictly unseen data the model never has access to during the training/validation phases of model creation, most closely resembles the data the model will be required to analyze in practice. Furthermore, as we endeavored to make our model as reproducible as possible by investigators at other institutions, we aimed to keep the model design and validation procedure simple.

In this case of binary classification, after a softmax is applied, the two output nodes (yes/no) have output values which add up to 1.0. If the positive node's output was greater than 0.5 (i.e., the output value of the "positive node" was greater than the "negative node" output), the prediction is treated as positive. Otherwise it was treated as a negative prediction. For sensitivity, 95% CI was calculated by treating each positively labeled example as a Bernoulli trial for the model—a positive prediction (TP) is a success, and a negative prediction (FN) is a failure; therefore the sensitivity is distributed binomially and can be approximated with standard methods for binomial proportion confidence intervals; we used the Wilson score interval. The inverse is done for positive predictive value.

4. Additional Model Performance Statistics

Validation set performance was recorded to assess model performance on unseen data. On the final epoch of the validation set, our network achieved an accuracy of 97.3% (95% confidence interval: 95.2–98.5%), a recall (sensitivity) of 100% (90.6–100%), a precision (positive predictive value) of 77.1% (63.5–86.7%), and an F1 score of 87.0%.

Appendix B Validation set result distribution

	Ground truth		
Model		+	−
	+	37	11
	−	0	361

Appendix C Test set result distribution

	Ground truth		
Model		+	−
	+	39	8
	−	3	359