Algorithmic Prediction of Delayed Radiology Turn-Around-Time during Non-Business Hours

Vaibhavi Shah, BS, Yeshwant R Chillakuru, MSc, Alex Rybkin, MD, Youngho Seo, PhD, Thienkhai Vu, MD, PhD, Jae Ho Sohn, MD, MS

Rationale and Objectives: Radiology turnaround time is an important quality measure that can impact hospital workflow and patient outcomes. We aimed to develop a machine learning model to predict delayed turnaround time during non-business hours and identify factors that contribute to this delay.

Materials and Methods: This retrospective study consisted of 15,117 CT cases from May 2018 to May 2019 during non-business hours at two hospital campuses after applying exclusion criteria. Of these 15,177 cases, 7,532 were inpatient cases and 7,585 were emergency cases. Order time, scan time, first communication by radiologist, free-text indications, and other clinical metadata were extracted. A combined XGBoost classifier and Random Forest natural language processing model was trained with 85% of the data and tested with 15% of the data. The model predicted two measures of delay: when the exam was ordered to first communication (total time) and when the scan was completed to first communication (interpretation time). The model was analyzed with the area under the curve (AUC) of receiver operating characteristic (ROC) and feature importance. Source code: https://bit.ly/2UrLiVJ

Results: The algorithm reached an AUC of 0.85, with a 95% confidence interval [0.83, 0.87], when predicting delays greater than 245 minutes for “total time” and 0.71, with a 95% confidence interval [0.68, 0.73], when predicting delays greater than 57 minutes for “interpretation time”. At our institution, CT scan description (e.g. “CTA chest pulmonary embolism protocol”), time of day, and year in training were more predictive features compared to body part, inpatient status, and hospital campus for both interpretation and total time delay.

Conclusion: This algorithm can be applied clinically when a physician is ordering the scan to reasonably predict delayed turnaround time. Such a model can be leveraged to identify factors associated with delays and emphasize areas for improvement to patient outcomes.

Key Words: Quality Improvement; Machine Learning; Work flow Efficiency; Natural Language Processing; Resident Education.

INTRODUCTION

The time taken from clinician ordering to radiology scan interpretation time is an important quality measure that can impact hospital workflow and patient outcomes. Delays in performing radiology scans or interpreting them can prolong the time until a proper treatment decision can be made, leading to increased costs and potential compromises in patient care (1). During non-business hours at our institution, the impact can be especially notable as a smaller number of radiologists is present, that often consists of radiology trainees or nighthawks instead of subspecialty attending radiologists. As a result, it can be more challenging to produce reports in a timely manner and cover a large number of time-sensitive studies while maintaining accurate interpretation, which may further contribute to delay (2).

Predicting whether a delay can occur after a radiological exam request can allow the ordering clinician to better prepare their immediate treatment plan. The clinicians can triage patients more efficiently and minimize the potential economic impact and patient care issues by accounting for the effect of delay time (3). Prediction of a delay after exam completion can inform radiologists on where possible setbacks may occur and how to rectify this in future training to create...
a better system of establishing turnaround times (4). For academic hospitals with residents taking overnight calls, this can also be an opportunity to identify any gaps in education and improve resident learning.

Machine learning can integrate various clinical features and clinician communication to result in a comprehensive prediction baseline, as seen in previous predictive studies (5). Binary classification is a commonly used application of supervised machine learning, and can be employed to distinguish between two expected outcomes (6). Natural language processing (NLP) in machine learning methods allows for the model to directly interface with text and language (7). However, while NLP has been recognized as a useful outlet, it is still a naive space that has yet to be fully tapped into, particularly for integrating free-form clinical text into predictive studies (8).

We aimed to develop and test a machine learning model aided by NLP components that could predict delay in interpretation of on-call radiology cases, followed by identification of the factors that contributed most to the prediction. This would allow other institutions to adopt and train similar models using our established approach. We hypothesized that the model could utilize a list of predictors such as inpatient status, the study description, body part, time of the day, PGY (post-graduate year) level in training, hospital campus, and written clinical history of the patient to predict whether a delayed turnaround time would occur.

MATERIALS AND METHODS

Data Acquisition

This was an institutional review board approved, consent-waived, single institution, multi-campus, retrospective analysis of CT diagnostic radiology reports from May 2018-April 2019 at a tertiary academic medical center with a large radiology training program. The original de-identified dataset consisted of 29,377 inpatient and emergency cases, acquired from the departmental radiology information system with the mPower search tool by the leading radiologist involved in this project.

Inpatient status (inpatient vs emergency), time of day (morning, afternoon, evening, late night), body part (Head/C-Spine, Extremity, Abdomen, Chest, Spine, Neck), study description (i.e. "CT chest pulmonary embolism protocol"), hospital campus (two distinct campuses), and PGY of radiologists interpreting the exam were collected as metadata. These six features were chosen by the authors as the most important factors available to a physician ordering a CT scan.

On-call, non-business hour cases were defined as those time-stamped on DICOM as completed during a weekend, holiday, or any other day between 17:00 and 08:00. Reports were excluded from the dataset if they were conducted during business hours (M-F, 08:00 to 17:00) (n = 9,608). Interventional radiology reports were excluded, image-only studies uploaded from outside hospitals were excluded, reports with missing PGY (attending physicians primarily read these cases) were excluded, and reports with other missing features were excluded (n = 4,578). Cases with trainee level listed as PGY-2 were excluded as they typically do not cover calls at our institution (n = 21).

The outcome variables, total time and interpretation time, were extracted from the radiology information system and free-text reports. The total time was defined as the first communication time minus clinician scan order time. Interpretation time was defined as the first communication time minus scan completion time. First communication time was defined as the earlier of preliminary report completion time (auto-generated variable) or verbal communication time (reported in the free-text). The free-text communication was extracted following symbols such as “/” which denoted communication occurrence at our institution; this was further verified with keywords such as "discussed" and "communicated".

The time was extracted with regular expressions first extracting all numbers and then identifying 'am', 'pm', or ':'. Any sentences that did not fit these criteria were manually labeled (n = 22). The extracted cases and times were manually confirmed to be correct for the first 200 reports. All cases with erroneously recorded times of communication in the free-text report (cases with a negative total/interpretation time) were excluded (n = 53). An exception to the exclusion was when interpretation time was within the negative 30-minute window, since these were almost always due to rounding errors (physicians chose a rounded-down completion time close to when the scan was completed) or communication happening during scan processing (communication before the scan was officially marked as complete by the system). These cases were set to interpretation time of 0 minutes to reflect quick communication (n = 364). These exclusions resulted in a final dataset of 15,117 cases (Fig 1).

Data Preprocessing

Categorical variables were one-hot-encoded: inpatient status (n = 2, inpatient or emergency), body part (n = 6 distinct divisions by subspecialty areas), study description (n = 101 distinct exam codes), time of day of the exam (n = 4, evening, late night, morning, and afternoon), PGY (n = 4: PGY 3, 4, 5 and 6; as ordered categorical variables), and hospital campus where exam was performed (n = 2, binary). PGY-6 includes all radiology fellows. One hot encoding of the categorical variables was achieved by making each categorical variable a binary column in the data table. “1” was used to denote the presence of the variable and “0” to note the absence.

Sentences were tokenized and timestamps were stored as datetime objects. CLEVER terminology was used to map medical terms as well as RadLex for radiology terms (9,10). Common modifications include substitution of the term "NEGEX" for “ruled out” and removal of irrelevant words and punctuation. Preprocessed text was tokenized into a sparse relative frequency matrix using CountVectorizer and TfidfTransformer (11,12). CountVectorizer was used to
generate a count matrix by assigning unique positions in a vector to build vocabulary. TfidfTransformer was used to reduce a bias towards preferring high frequency words that may cause the model to overfit to unimportant words. All preprocessing steps were conducted in Python 3.6.2 with the NLTK and sklearn package (11,12,13).

**Model Training**

The final dataset consisting of 15,117 cases were split into training/validation for 5-fold cross-validation and testing sets with 85% (12,849) and 15% (2,268) of the data, respectively. An ensemble approach was used to integrate a classification and natural language processing model (Fig 2) (14). An XGBoost classification model was used on the aforementioned metadata (15). A Bag of Words model was used in conjunction with a Random Forest for the free-text (16). The one-hot-encoded categorical variables and the sparse matrix representation of free-text were inputted into the XGBoost and Bag of Words model, respectively. Hyperparameters (including ensemble weights) were chosen using Stratified 5-Fold Cross-Validation on the training set, employing a Grid Search method. XGBoost hyperparameters were 100 estimators (trees used), a maximum depth of 100, and a learning rate of 0.1.

**Figure 1.** Criterion for inclusion/exclusion of radiology reports for dataset. Interventional reports, image-only outside reports, reports with missing data, reports with erroneously recorded times, and reports covered by PGY-2 (post-graduate year) trainees were excluded from the on-call CT (computed tomography) scans to obtain the final dataset. (Color version of figure is available online.)

**Figure 2.** Natural language processing and machine learning model for delay prediction. Raw clinical history text was preprocessed with CLEVER and Radlex libraries to encode meaningful words into vectors. These vectors were used in training the Random Forest Classifier. Metadata features were numerically labelled and used in the XGBoost Classification model. Both models’ predictions were integrated in an ensemble approach to obtain a final prediction. (Color version of figure is available online.)
5 in each decision tree, and a learning rate of 0.5. Random Forest hyperparameters included 100 trees, no maximum depth, minimum of 16 samples per node, and 1 data point per leaf. XGBoost and Random Forest outputs were evenly weighted to generate the final ensemble prediction for whether a case would be delayed.

Delay was defined as greater than 245 minutes for total time and greater than 57 minutes for interpretation time. These thresholds were set according to the mean of the total and interpretation times of the dataset (Table 1).

Model Evaluation
The model was tested on the 15% hold-out from the dataset. Probability that the report was indicative of delay was obtained for the features and clinical text alone, and then weighted to obtain an overall probability for classification. A Receiver Operating Characteristic (ROC) curve for the dataset was obtained with a corresponding Area Under the Curve (AUC) score. Features were also plotted based on importance in the XGBoost to identify the most relevant variables for prediction.

The XGBoost classifier of radiology report metadata indicated feature importance, also known as Gini importance or Mean Decrease in Impurity, of the input parameters. Feature importance is a feature of XGBoost that measures the relative ability of each feature to improve model accuracy at branch-points of tree-based models. A value closer to 1 on a scale from 0 to 1 reflects a more important feature in the decision making process of the XGBoost classifier. The most important words were extracted from the Bag of Words vocabulary by filtering for those with the highest frequencies. Full source code is made available at: https://bit.ly/2UrLiVJ

Statistical Analysis
For significance testing in exploratory data analysis, Wilcoxon Rank Sum Test for nonparametric distributions was conducted (Inpatient Status, Hospital Campus). Kruskal Wallis test for non-parametric distributions was applied for categories with greater than 2 groups to account for multiplicity (Time of Day, Day of the Week, Body Part, PGY) and determine differences in means. All analyses were conducted in the SciPy library of Python, with a two-sided test and level of significance at 0.05.

RESULTS
Demographics
The dataset was composed of 15,117 CT reports (Table 1). 7,532 cases were defined as inpatient cases, while 7,585 were emergency cases. 14,169 cases were conducted at Hospital Campus 1, and 948 cases were conducted at Hospital Campus 2. Campus 1 is an adult tertiary referral center that sees a larger number of medical and surgical patients, whereas Campus 2 is a smaller campus specializing in pediatric, obstetrical, and cancer care, accounting for the difference in case volume between the two locations. There were a majority of abdomen, Head/C-Spine, Chest CT scans while there were not as many Spine, Neck, and Extremity scans. Mean turnaround times for these categories are also shown in Table 1.

The average of the total time delay was 4.08 hours ± 0.03 hours, approximately 245 minutes. The average of the interpretation time delay was 56.61 ± 0.77 minutes.

Model Evaluation
The ROC curves of the ensembled model were trained on 85% of radiology reports and tested on the remaining 15%

| Criteria              | Cases   | Total Time (hours) | Interpretation Time (mins) |
|-----------------------|---------|--------------------|---------------------------|
| Patient Status        |         |                    |                           |
| Inpatient             | 7532    | 6.00 ± 0.06        | 69.38 ± 1.31*             |
| Emergency             | 7585    | 2.18 ± 0.02        | 43.92 ± 0.80*             |
| Point of Care         |         |                    |                           |
| Hospital 1            | 14169   | 4.10 ± 0.04        | 56.48 ± 0.80              |
| Hospital 2            | 948     | 3.87 ± 0.11        | 58.43 ± 3.35              |
| Time of Day           |         |                    |                           |
| Morning               | 1239    | 4.15 ± 0.14        | 59.25 ± 3.20*             |
| Afternoon             | 1608    | 4.11 ± 0.10        | 63.43 ± 2.59*             |
| Evening               | 6290    | 4.11 ± 0.05        | 55.27 ± 1.15*             |
| Late Night            | 5980    | 4.03 ± 0.05        | 55.63 ± 1.21*             |
| Day of the Week       |         |                    |                           |
| Monday                | 1819    | 4.02 ± 0.09        | 57.47 ± 2.29              |
| Tuesday               | 1993    | 4.13 ± 0.09        | 54.38 ± 1.97              |
| Wednesday             | 2001    | 4.22 ± 0.09        | 56.40 ± 2.24              |
| Thursday              | 1880    | 4.14 ± 0.11        | 52.35 ± 1.99              |
| Friday                | 1865    | 4.18 ± 0.11        | 52.22 ± 2.00              |
| Saturday              | 2757    | 4.14 ± 0.08        | 61.42 ± 1.86              |
| Sunday                | 2802    | 3.84 ± 0.07        | 58.82 ± 1.96              |
| Body Part             |         |                    |                           |
| Abdomen               | 4742    | 4.12 ± 0.06        | 43.65 ± 1.04*             |
| Chest                 | 3100    | 4.97 ± 0.09        | 46.18 ± 1.32*             |
| Head/C-Spine          | 5517    | 3.52 ± 0.05        | 64.72 ± 1.38*             |
| Extremity             | 404     | 4.15 ± 0.21        | 51.47 ± 4.64*             |
| Spine                 | 1024    | 3.71 ± 0.13        | 91.15 ± 4.53*             |
| Neck                  | 330     | 5.71 ± 0.31        | 104.35 ± 9.62*            |
| PGY Level             |         |                    |                           |
| 3                     | 3667    | 3.41 ± 0.06        | 42.47 ± 0.96*             |
| 4                     | 5900    | 3.74 ± 0.05        | 48.45 ± 1.00*             |
| 5                     | 751     | 4.08 ± 0.16        | 27.00 ± 1.34*             |
| 6                     | 4799    | 5.03 ± 0.07        | 82.07 ± 1.91*             |
| All Cases             | 15117   | 4.08 ± 0.03        | 56.61 ± 0.77              |
are shown in Figure 3. The AUC for prediction of interpretation time was 0.71, with a 95% confidence interval [0.68, 0.73]. The AUC for prediction of total time was 0.85, with a 95% confidence interval [0.83, 0.87].

For interpretation time delays, when the model was set to a sensitivity of 85% (513 out of 604), for example, it achieved a specificity of 39% (648 out of 1664). The Positive Predictive Value (PPV, Precision) was 0.34, the Negative Predictive Value (NPV) was 0.88, the F1 Score was 0.48, and the MCC (Matthews Correlation Coefficient) was 0.23. For total time delays, when the model was set to a sensitivity of 52% (394 out of 751), for example, it achieved a specificity of 91% (1380 out of 1517). The Positive Predictive Value (PPV, Precision) was 0.74, the Negative Predictive Value (NPV) was 0.79, the F1 Score was 0.61, and the MCC (Matthews Correlation Coefficient) was 0.48.

Of note, a separate set of ensemble models was also evaluated that only differed by using individual radiology trainees, rather than PGY, as input variables. These models had a performance (AUC score) of 0.81 (95% confidence interval [0.79, 0.83]) and 0.85 (95% confidence interval [0.84, 0.87]) for interpretation time and total time delay, respectively (Fig S3).

**Feature Importance**

The study description, time of day, and on-call radiologist’s PGY were the most important factors of the model (Fig 4). The study description had feature importance values of 0.43 and 0.49 for interpretation time delay and total time delay, respectively. Time of day had feature importance values of 0.18 and 0.16, and the on-call radiologist PGY demonstrated feature importance values of 0.18 and 0.17 for interpretation time delay and total time delay, respectively. Notably, feature importance on a separate model that only differed in utilizing individual trainees, rather than PGY, demonstrated that radiologist had the highest importance (Fig S3), suggesting that

---

**Figure 3.** Receiver operating characteristic (ROC) curves of ensembled tree-based classifier models trained on 85% of on-call radiology reports and tested on the remaining 15% of reports. The ROC curve labeled “Interpretation time” represents reporting delays of greater than 57 minutes in the time from when the exam is completed to when the first communication/preliminary report is made. The ROC-AUC score for interpretation time is 0.71, with a 95% confidence interval [0.68, 0.73]. The ROC curve labeled “Total time” represents reporting delays of greater than 245 minutes in the time from when the exam is ordered to when the first communication/preliminary report is made. The ROC-AUC score for total time is 0.85, with a 95% confidence interval [0.83, 0.87]. (Color version of figure is available online.)

**Figure 4.** Feature importance values of XGBoost Classifier trained on 85% of radiology report metadata for predicting delays in both “Interpretation Time” and “Total Time.” The on-call radiologist PGY (post-graduate year) had feature importance values of 0.18 and 0.17 for interpretation time and total time delay, respectively. Time of day had feature importance values of 0.18 and 0.16 for interpretation time and total time delay, respectively. The study description had feature importance values of 0.43 and 0.49 for interpretation time and total time delay, respectively. (Color version of figure is available online.)
individual factors were more influential than level of training on turnaround time.

While the six features used in the XGBoost model in this paper were chosen directly by the authors, an additional analysis utilizing a feature selection process is included in the Supplement (Fig S4 and Fig S5). This feature selection was done using the XGBoost feature importance scores to reduce the number of features to the top three.

The text integration was also indicative of delay prediction as shown by the most relevant keywords to the model. Indicative words for prediction of both total time and interpretation time delay included “fall”, “pain”, and “screen”.

Error Analysis

Demonstrated in Table 2 and 3 are representative cases for both correct and incorrect predictions of delay. In these tables, confidence is derived from the predicted probability of the model (the probability the model will predict a certain case is delayed). A value closer to 1 reflects a more confident prediction. In terms of error, false positives (algorithm predicted delay but there was no delay) were common in cases that involved routine cancer screening and pre/post-operative checks. These were generally deprioritized at our institution, but these instances could occasionally get more reprioritized if acute abnormalities were found. False negatives (algorithm predicted no delay but there was delay) were common in trauma, altered mental status cases, and cases meant to rule out acute appendicitis or renal failure. These cases typically have a rapid turnaround time but if there was complex finding then an occasional delay could have occurred.

Manual inspection of the error cases by a senior radiology resident (PGY-5) revealed no other discernible systematic error pattern on the false positive cases, but revealed that in many false negative cases, the indications were for non-emergent studies that likely received lower priority by the overnight radiologist when there were other emergent studies to interpret.

DISCUSSION

We have developed an ensemble machine learning tool with a natural language processing component that can predict delays in radiology services during non-business hours.

As delays in radiology are an important measure of patient safety and hospital efficiency, having the ability to predict such potential delays has important benefits. Furthermore, prediction of delays in radiology can improve the referrer and radiologist relationship and help clinicians to prepare alternative...
options in case a delay is expected. Especially during non-business hours, when less senior-level radiologists are present and studies are generally more emergent than during business hours, characterizing this delay can be especially useful. Most importantly, such a prediction algorithm can be a starting point for quality improvement projects that can reduce delays in radiology turnaround time during the unique non-business hour workflow. Given the complexity of real-world radiology workflow, no algorithm can make perfect predictions on which cases will be delayed. However, attaining a reasonable prediction of such cases can be relevant.

At our institution, the feature importance of CT study description, and to some extent, PGY and time of day, were multiple fold greater than importance values for body part, inpatient status, or hospital campus. These trends in feature importances were similar for both interpretation and total time, likely due to their interdependence (“interpretation time” plus time taken to complete the scan yields “total time”). Inherently, the presence of trainees ranging from PGY-3 to PGY-5 and fellow level (labeled PGY-6), in addition to individual ability even at a specific PGY level, likely contributed to the heterogeneity in turnaround time. Of note, when evaluating a similar model built using individual trainees, rather than PGY, interpretation time delay predictions improved, ROC-AUC 0.71 increased to ROC-AUC of 0.81, but not total time delay predictions—ROC-AUC stayed at 0.85 (Fig S2). This suggests that even within a PGY level, variance in individual trainees can contribute notably to delays.

Additionally, certain protocols tended to be associated with delayed turnaround time due to the inherent complexity of the cases, most notably CT neck with contrast and CTA aortic dissection protocol that often spanned multiple body parts. Similarly, the relative importance of time of day may suggest that this variable is capturing variation in relative workload volumes throughout the on-call schedule, such as high case volume during weekday evening rush hours.

Error analysis by a radiologist revealed several erroneous predictions of delay (false positive) in cancer screening and non-emergency pre-operative planning scans. This is an expected error because most routine imaging indications requested to be completed overnight receive lower priority. Occasionally, some of these scans get completed quickly and interpreted with no delay if the hospital is not busy or the radiologist identified an unexpected acute finding. Another notable category of error was erroneously predicted no delay cases (false negative), which was most commonly seen in routine CT brain without contrast scans. Most brain CT scans were done for trauma at our institution and were read with quick turnaround time due to the likelihood of negative scans.

### TABLE 3. Error Analysis for Prediction of Interpretation Time Delay

The top 5 representative cases for correct, confident and incorrect, confident predictions are shown for total time delay greater than 57 minutes. Confidence reflects the probability the model predicts delay or no delay for a given case. Inpatient is abbreviated IP and Emergency is abbreviated E.

| Algorithm Predicts Delay Correctly | Features | Confidence |
|-----------------------------------|----------|------------|
| **Clinical History Transformed**  |          |            |
| Cervicothoracic laminectomy, fusion, prosthetic infection | IP, Spine, Evening | 0.777 |
| Right middle cerebral artery infarct, trisomy | IP, Neck, Late Night | 0.736 |
| Left leg weakness, C3-T1 posterior spinal fusion | IP, Spine, Late Night | 0.665 |
| Altered mental status, neck pain | E, Head/C-Spine, Evening | 0.642 |
| Motorcycle accident | E, Spine, Late Night | 0.628 |

| Algorithm Predicts No Delay Correctly | Features | Confidence |
|--------------------------------------|----------|------------|
| **Clinical History Text Transformed** |          |            |
| NEGEX_renal calculus | E, Abdomen, Evening | 0.974 |
| Two long bone fracture, SCREEN ICH | E, Head/C-Spine, Evening | 0.964 |
| NEGEX_subarachnoid hemorrhage | E, Head/C-Spine, Evening | 0.955 |
| SOB hypoxia | E, Chest, Evening | 0.946 |
| Abdominal pain, r/o appendicitis | E, Abdomen, Late Night | 0.925 |

| Algorithm Predicts Delay Incorrectly | Features | Confidence |
|-------------------------------------|----------|------------|
| **Clinical History Text Transformed** |          |            |
| Preoperative | IP, Spine, Late Night | 0.690 |
| Status post aneurysm clipping | IP, Head/C-Spine, Late Night | 0.641 |
| End-stage renal disease status post basal transplant | IP, Neck, Morning | 0.627 |
| Neutropenic fevers, B-cell acute lymphoblastic leukemia | IP, Head/C-Spine, Morning | 0.618 |
| Preoperative CT angiogram | IP, Head/C-Spine, Late Night | 0.612 |

| Algorithm Predicts No Delay Incorrectly | Features | Confidence |
|----------------------------------------|----------|------------|
| **Clinical History Text Transformed**  |          |            |
| NEGEX_appendicitis | E, Abdomen, Morning | 0.978 |
| NEGEX_renal_calculus | E, Abdomen, Evening | 0.937 |
| Fall ICH | E, Abdomen, Late Night | 0.914 |
| Left sided pain, fever | IP, Abdomen, Evening | 0.906 |
| SCREEN_pulmonary embolism | E, Chest, Evening | 0.897 |
and potential seriousness of delayed interpretation in the chance of acute intracranial findings. Occasionally, there were complex trauma cases that took up a lot of time or follow up CT of the brain that received lower priority, which likely accounted for the error.

It is important to note that the primary objective of the model is to predict a future event (of delayed turnaround time) and there is inherent uncertainty present. The model should be considered a learned representation of the patterns from the past data that may predict the future rather than a definitive answer. Arguably, the more important value of the model is that it can be an important measure of past quality in the radiology workflow and an opportunity to reveal areas that may benefit from a formal improvement process.

However, the algorithm can be useful for clinicians to get an estimation on whether a report would be delayed or not based on the institution's past data. This can allow the physician to gauge when to expect results and handle care in the intermittent time. Clinicians can prepare for the delay by better triaging patients and seeking alternative diagnostic approaches if necessary. It can also be used on the “back end” of operations by departments that want to improve turnaround times, identify any associations with delays, and how they can be avoided.

The study had several limitations. First, it is a single institution study and generalizability of our institution specific weights remain indeterminate. However, an important point of this study is to present a viable and verified approach that can be undertaken to develop a delay prediction model at each institution. Additionally, we fully released the source code so the study can be replicated at other institutions. Second, MRI and ultrasound cases were not included in the study because at our institution these were not performed overnight unless explicitly requested by clinicians to call in a technologist, leading to variable and unpredictable delays. Third, the vocabulary of the NLP model is reliant on the premade Radlex and CLEVER libraries, which limits the ability of the model to account for newly emerging phrases in medicine and institution-specific jargon. These can be manually updated to ensure the best outcome of the model. Fourth, there were additional variables such as transportation times and technologist factors that could have further improved the model's accuracy but could not be collected for this study. Fifth, our model was unable to explicitly account for the small number of cases (<5%) that were deprioritized overnight (e.g. cancer screening) since this was largely dependent on each radiologist's discretion.

Future study considerations include collecting additional variables such as hospital census, emergency department patient census, number of providers, or transportation and average technology operating time which were not readily available to us at the time of the study. Additionally, a follow up full quality improvement project that completes the PDSA (plan-do-study-act) cycle could be a next step of the project. After clinical integration of the algorithm, further feedback from the clinical community would be important to identify any preventable pitfalls and carefully assess any potential harms that may arise from false positive or false negative errors committed by the model.

CONCLUSION

We have developed an ensemble machine learning model that can predict delayed radiological turnaround time during non-business hours and proposed an approach to identify factors that contribute to the delay. Such a model can be leveraged to identify factors associated with delays and emphasize areas for improvement to patient outcomes.

ABBREVIATIONS

CT - Computed Tomography (Definition according to the National Cancer Institute, “a procedure that uses a computer linked to an x-ray machine to make a series of detailed pictures of areas inside the body”)

ED - Emergency Department (Definition according to MedicineNet, “the department… responsible for the provision of medical and surgical care to patients arriving at the hospital in need of immediate care”)

NLP - Natural Language Processing (Definition according to Machine Learning Mastery, “automatic manipulation of natural language, like speech and text, by software”)

PGY - Post-Graduate Year (Definition according to “An Introduction to ROC Analysis” by Tom Fawcett, “the probability that a classifier will rank a positive instance higher than a negative instance”)

FUNDING AND GRANT

Sponsors has no role in the study design, implementation, or publication submission. Sponsors has no role in the study design, implementation, or publication submission. The author(s) declare(s) that they had full access to all of the data in this study and the author(s) take(s) complete responsibility for the integrity of the data and the accuracy of the data analysis. The authors state no conflict of interest or competing interests. JHS was supported by the NIBIB T32-EB001631.

KEY FINDING

We developed a machine learning and natural language processing tool that predicts delays in radiology services during non-business hours. The algorithm reached an AUC of 0.85, with a 95% confidence interval [0.83, 0.87], when predicting “total time” delays and 0.71, with a 95% confidence interval [0.68, 0.73], for “interpretation time” delays. At our institution, CT study description was the most predictive feature.
IMPORTANCE

This algorithm can be applied when a physician is ordering a CT to estimate whether the report will be delayed based on clinical features.

DECLARATIONS OF INTEREST

None

REFERENCES

1. Cournane S, Conway R, Creagh D, Byrne DG, Sheehy N, Silke B. Radiology imaging delays as independent predictors of length of hospital stay for emergency medical admissions. Clin Radiol 2016; 71(9):912–918.
2. Little D, McCoubrie P. Learning whilst on-call: a vital part of radiology training? Clin Radiol 2016; 71(9):921–924.
3. Imperato J, Morris DS, Binder D, et al. Physician in triage improves emergency department patient throughput. Intern Emerg Med 2012; 7(5):457–462.
4. England E, Collins J, White RD, Seagull FJ, Deledda J. Radiology report turnaround time: effect on resident education. Acad Radiol 2015; 22(5):662–667.
5. Wang S, Summers RM. Machine learning and radiology. Med Image Anal 2012; 16(5):933–951.
6. Maglogiannis IG. Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies. IOS Press, 2007.
7. Kim Y. Convolutional Neural Networks for Sentence Classification. Proc 2014 Conf Empir Methods Nat Lang Process EMNLP. Doha, Qatar: Association for Computational Linguistics; 2014, p. 1746–1751. Available at: https://www.aclweb.org/anthology/D14-1181. Accessed August 1, 2019.
8. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Medical Informatics 2019; 7(2). doi:10.2196/12239.
9. RadLex Term Browser. Available at: http://www.radlex.org/. Accessed July 30, 2019.
10. Banerjee I, Li K, Seneviratne M, et al. Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. JAMA Open 2019; 2(1):150–159.
11. sklearn.feature_extraction.text.CountVectorizer — scikit-learn documentation. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html. Accessed May 4, 2020.
12. sklearn.feature_extraction.text.TfidfVectorizer — scikit-learn documentation. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html. Accessed May 4, 2020.
13. nltk Package — NLTK 3.5 documentation. Available at: https://www.nltk.org/api/nltk.html. Accessed May 5, 2020.
14. Lutins E. Ensemble Methods in Machine Learning: What are They and Why Use Them? Medium 2017. Available at: https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9f6f5f. Accessed July 30, 2019.
15. Understand your dataset with XGBoost — xgboost 1.0.0-SNAPSHOT documentation. Available at: https://xgboost.readthedocs.io/en/latest/R-package/discoverYourData.html. Accessed July 30, 2019.
16. Random Forest in Python - Towards Data Science. Available at: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0. Accessed July 30, 2019.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.acra.2021.05.026.