Journal of Digital Imaging
https://doi.org/10.1007/s10278-019-00260-w

ORIGINAL PAPER

Determining Follow-Up Imaging Study Using Radiology Reports

Sandeep Dalal[1] · Vadiraj Hombal[1] · Wei-Hung Weng[2] · Gabe Mankovich[1] · Thusitha Mabotuwana[3] · Christopher S. Hall[3] · Joseph Fuller III[4] · Bruce E. Lehnert[4] · Martin L. Gunn[4]

Abstract
Radiology reports often contain follow-up imaging recommendations. Failure to comply with these recommendations in a timely manner can lead to delayed treatment, poor patient outcomes, complications, unnecessary testing, lost revenue, and legal liability. The objective of this study was to develop a scalable approach to automatically identify the completion of a follow-up imaging study recommended by a radiologist in a preceding report. We selected imaging-reports containing 559 follow-up imaging recommendations and all subsequent reports from a multi-hospital academic practice. Three radiologists identified appropriate follow-up examinations among the subsequent reports for the same patient, if any, to establish a ground-truth dataset. We then trained an Extremely Randomized Trees that uses recommendation attributes, study meta-data and text similarity of the radiology reports to determine the most likely follow-up examination for a preceding recommendation. Pairwise inter-annotator F-score ranged from 0.853 to 0.868; the corresponding F-score of the classifier in identifying follow-up exams was 0.807. Our study describes a methodology to automatically determine the most likely follow-up exam after a follow-up imaging recommendation. The accuracy of the algorithm suggests that automated methods can be integrated into a follow-up management application to improve adherence to follow-up imaging recommendations. Radiology administrators could use such a system to monitor follow-up compliance rates and proactively send reminders to primary care providers and/or patients to improve adherence.

Keywords Medical informatics applications · Radiology · Natural language processing · Supervised machine learning · Follow-up studies

Introduction

Diagnostic radiologists interpret medical images for a patient to identify and characterize disease, quantify biomarkers of the disease severity, and monitor disease progression or response to therapy. Their primary work product is the radiology report, which is communicated to the referring clinician and subsequently to the patient.

✉ Vadiraj Hombal
vadiraj.hombal@philips.com

Sandeep Dalal
sandeep.dalal@philips.com

Wei-Hung Weng
ckbjimmy@gmail.com

Gabe Mankovich
gabe.mankovich@philips.com

Thusitha Mabotuwana
thusitha.mabotuwana@philips.com

Christopher S. Hall
cshall2015@gmail.com

Joseph Fuller, III
joseph.trey.fuller@gmail.com

Bruce E. Lehnert
blhnrt2@gmail.com

Martin L. Gunn
marting@uw.edu

[1] Clinical Informatics Solutions and Services, Philips Research North America, 2 Canal Park, Cambridge, MA 02141, USA

[2] Department of Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, MA 02115, USA

[3] Radiology Solutions, Philips Healthcare, 22100 Bothell Everett Highway, Bothell, WA 98021, USA

[4] Department of Radiology, University of Washington, 1959 NE Pacific Street, Seattle, WA 98195, USA

Published online: 26 August 2019

Approximately 6–12% of radiology reports contain follow-up recommendation for further action such as follow-up imaging studies [1]. Although radiologists are responsible for conveying this information to the referring clinician, the referrer is usually responsible for conveying this information to the patient and ensuring the patient schedules or undergoes the necessary study, where clinically appropriate [2].

With the exception of specific screening programs (e.g., mammography), radiology departments often do not have a means to automatically track which patients have been recommended for follow-up imaging and, more significantly, may not know if patients have scheduled and completed the appropriate follow-up imaging study in a timely manner. Even when the patient has completed the follow-up imaging study, this information is not typically explicitly recorded in any of the hospital systems, such as the electronic health record or radiology information system.

In one recent study, over one-third of all recommendations were not followed-up [3]. Of those not followed-up, 40% of recommendations were not acknowledged by the referring clinician and hence not followed up; 44% of those patients were at risk of significant harm—e.g., suspected cancer. Similarly, another study examining incidental pulmonary nodules found a low follow-up rate of 29% [4]. Kulon et al. describe a system that performs the detection, filtering, sorting, and management of recommendations and facilitates the secure communication of such recommendations with clinicians or patients [5]. Sloan et al. found that approximately 12% of patients did not receive follow-up care for potential cancer within 3 months of the radiology recommendation, primarily due to shortcomings in the means of electronic communication [6]. Cook et al. describe an automated recommendation-tracking engine to identify and monitor imaging follow-up among patients with liver, kidney, pancreatic, and adrenal lesions [7]. All prior efforts have focused on specific patient populations, conditions, or modalities, and as such, it is important to develop scalable methods to reliably identify patients who need to be followed-up from all radiology reports. This research is motivated by the practical problem of managing the large number of patients with follow-up recommendations. This process is largely manual today and as a result the cost is sufficiently high to only allow select patient populations to realistically be managed.

To address some of these existing limitations, we developed natural language processing and machine learning–based algorithms that can reliably determine the most likely follow-up examination, if any, from a given list of candidate radiology reports for the same patient. Such algorithms can be incorporated into routine practice to reduce the rate of patients lost to follow-up and to ensure that more patients have appropriate follow-up in a timely manner.

Methods

A system for automatic retrospective tracking and auditing of follow-up recommendations is composed of two sub-systems: (1) Automatic detection of recommendations and (2) automatic matching of the original examination recommendation to the most likely follow-up exam which satisfies the clinical reason(s) for follow-up.

Methods to extract recommendations from radiology reports have been reported in literature [8–11]. In previous work, we developed natural language processing methods to automatically extract recommendation sentences [12–14]. The corresponding recommendation-related attributes such as follow-up time interval, and modality were also extracted. Using 532 reports annotated by three radiologists (all authors in present study) as the ground truth, the detection algorithm was evaluated to have 97.9% accuracy.

In this work, we focus on determining the most likely follow-up exam, if any, among a set of subsequent radiology reports for the same patient. Although the detection algorithm distinguishes between nine different types of recommendations in radiology reports (e.g., imaging recommendations, clinical or therapy follow-up, tissue sampling or biopsy, and so on), the focus of the current work is on follow-up imaging recommendations only.

Data

This study was approved by the local institutional review board. We used a database of radiology reports generated between 2010 and 2013 from three network hospitals stored in the radiology information system of the Department of Radiology. Using the recommendation detection algorithm, we selected 564 imaging-studies containing follow-up imaging recommendations found in the Findings or Impression sections of reports. The studies containing the follow-up recommendations were selected to cover the common imaging modalities proportionate to their occurrence in the larger database; however, we excluded mammography studies as their follow-up is well prescribed. A summary of the statistics of the recommendation attributes for the selected studies is shown in Table 1. Note that in Table 1(a), a single recommendation can contain more than one imaging modality (e.g., "follow-up CT recommended in 3 months and MR in 6-12 months"), and "recommended modality" is the modality associated with the suggested follow-up exam.

Table 1 Number of studies with recommended modality and interval

	Number of studies
(A) Recommended modality	
CR/computed radiography	88
CT/computed tomography	245
PT/Position emission tomography	12
MG/mammography	12
MR/magnetic resonance	157
US/ultrasound	117
NM/nuclear medicine	12
XA/X-ray angiography	1
RF/radio fluoroscopy	1
OT/other	1
(B) Recommended interval	
Same day	18
2–6 days	44
7–14 days	34
15–30 days	38
1–3 months	119
3–6 months	39
6–12 months	54
Not explicitly specified	218

Annotation of Follow-up Exam

The same three radiologists who previously annotated follow-up recommendations reviewed the recommendation sentences and annotated the recommended imaging modality and appropriate follow-up timeframe. Five recommendations identified by the algorithm were identified by the radiologists as not being true follow-up imaging recommendations. Excluding these recommendations reduced the number of source reports with recommendations from 564 to 559.

There were 8691 subsequent reports corresponding to these 559 source exams, with a median of 11 candidate exams per source exam. The distribution of candidate exams is shown in Fig. 1.

For each recommendation, the three radiologists then annotated the most appropriate candidate follow-up examination if one existed. An analysis of the annotations by the three-radiologists is shown in Table 2. The Cohen's kappa statistic that measures pairwise inter-annotator agreement ranged from 0.744 to 0.792. At least two radiologists marked the same follow-up examination as the true follow-up examination in 387 cases, and in 172 cases at least two radiologists marked no examination as the true follow-up examination.

The recommended modality annotations from the three radiologists were reconciled by selecting the modality that was in agreement with at least two radiologists' annotations. The recommended time-frame was reconciled by selecting the median interval from the time-frames of all radiologists arranged in a temporal order. This reconciled dataset was used as the ground-truth for follow-up matching algorithm development and performance evaluation.

Algorithm Development

We used a two-stage approach for identifying the relevant follow-up examination. Stage 1 is used to compute the probability of a candidate examination being the follow-up to the preceding source recommendation sentence, and to select the likely candidates for a follow-up from the set of all candidate studies. For each candidate exam, a probability of that candidate being the follow-up exam is estimated using the extremely randomized trees (ERT) [15]. If all the candidates have a probability below a specified threshold, the recommendation is determined to not have been followed-up. Stage 2 then ranks each candidate exam based on its probability and time difference between source and candidate to determine the most relevant follow-up candidate.

This two-stage method provides both a score for all follow-up candidate reports and the ability to select one among them as the follow-up study. The classifier scores each candidate follow-up examination with a probability between 0 and 1. All candidates with a score greater than 0.5 are considered likely candidates. We used the heuristic that among the set of likely studies, the earliest study with a score within 95% of the maximum score is selected as the follow-up study.

Report Processing Stage 1 classifier uses features based on the patient, study meta-data and the report text of both the source and candidate reports. We employ a common report processing pipeline (Fig. 2) to extract features based on the report text.

The section parser segments the semi-structured raw report text into section header and content pairs. It further classifies the section header according to their semantic functions and assigns each section header with one of the following semantic labels: procedure, indication, history, comparison, findings, impression, addendum, correction, consent, and unknown.

In the next step, the text in the sections is first normalized by standardizing dates and times to a single format before tokens and part of speech tags are identified. The chunker then uses the tags to identify noun phrases (NP) in the text (Fig. 3).

Anatomy Detection The anatomy detection module is based on a dictionary of anatomies created using the

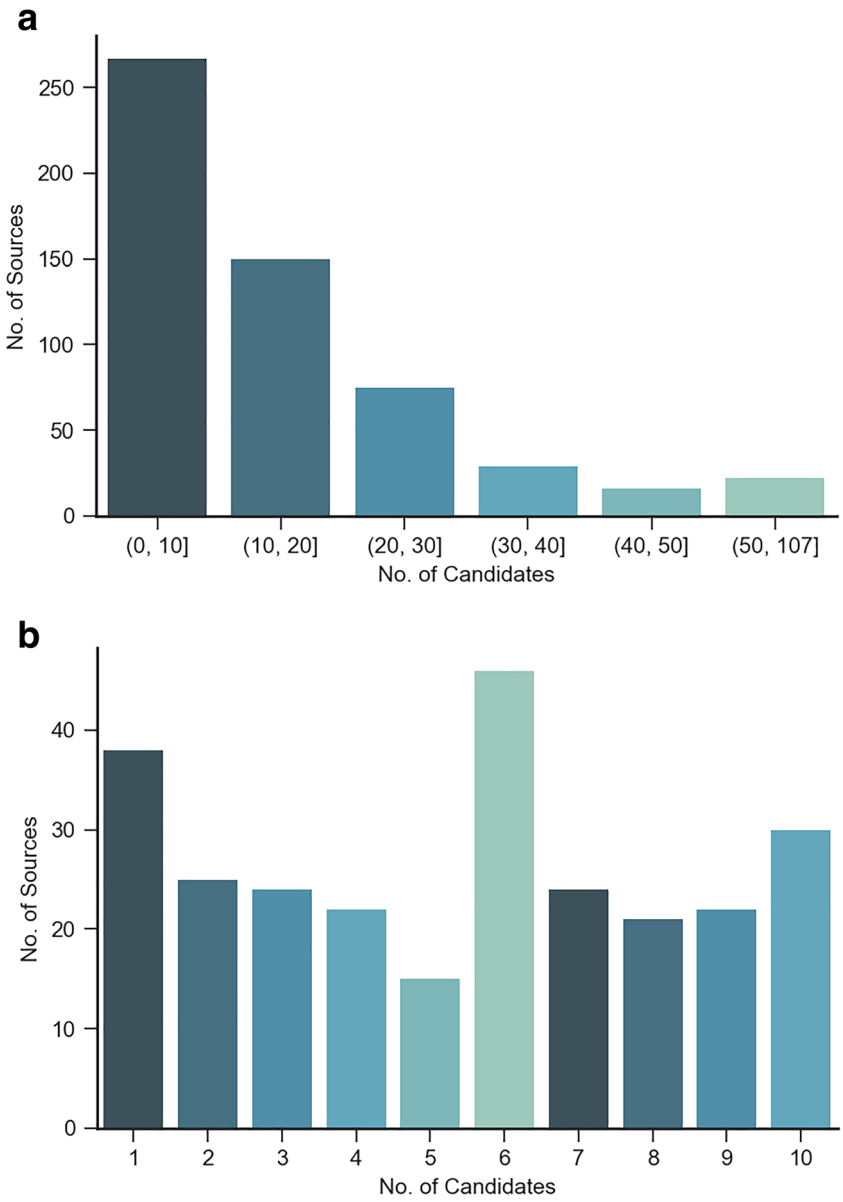

Fig. 1 Distribution of source exams **a** distribution of the entire data **b** source distribution with (0, 10) candidates

"subdivision of cardinal body part" and "organ" hierarchies of the Foundational Model of Anatomy (FMA) Ontology [16]. This module identifies anatomical phrases from raw text and maps them to 82 anatomical categories such as head, face, nasal, or orbit eye, which are relevant to matching content in radiology reports. This allows for phrases such as "right upper lobe of lung" and "pulmonary nodules" to be mapped to high-level anatomical regions such as "chest" and "lung."

Table 2 Follow-up concordance among annotators

# Annotators marking true or no follow-ups	# Source accessions
3 Annotators marked the same follow-up candidate	304
2 Annotators marked the same follow-up candidate	83
2 Annotators marked no follow-up candidate	41
3 Annotators marked no follow-up candidate	131
Total	559

The anatomy detection module was validated against 1135 sentences that were categorized for anatomies by a trained medical doctor. It achieved a precision score of 0.950, recall of 0.915, specificity of 0.842, and F_1 measurement of 0.932 against ground truth. Among 750 sentences with the annotated radiology-related anatomical categories, the module identified the anatomical categories completely correct in 713 (95.0%) sentences, partially correct or incorrect in 22 (2.9%)

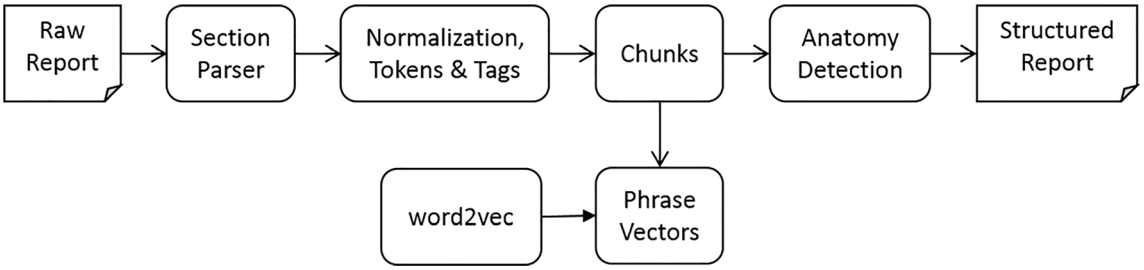

Fig. 2 Report processing pipeline

sentences and did not identify any anatomical categories in 15 (2.0%) sentences with annotated anatomical categories Table 3.

Phrase-Based Text Similarity The report text of a follow-up examination often addresses the context of the recommendation. Consequently, a high degree of semantic similarity may be expected between the two related reports. For recommendation and corresponding follow-up example shown in Fig. 4, the key to establishing relatedness is the ability to analyze phrases such as "bilateral subcentimeter pulmonary nodules" and "multiple bilateral lung nodules" and determine that they are referring to the same finding (i.e., semantically similar).

We employ a noun-phrase (NP) based text matching methodology to estimate semantic similarity between text spans based on a neural-network generated distribution of vector representations of words that captures the semantic and syntactic relationships between the words [17, 18]. The neural-network (the Skip-gram model with a 5-word window-size and negative sampling) was trained on a corpus of 1.6 million sentences from radiology reports and contains a vocabulary of 15,980 words. Specifically, we are interested in estimating similarity $S(T_1, T_2)$ between two texts $T_1$and T_2 which may be of any span length—sentences, paragraphs, or entire reports. Each text is represented only by the NPs in it. Therefore, $S(T_1, T_2) \approx S(H_1, H_2)$, where each H_1 and H_2 represent the NPs in the corresponding text. Specifically, $H_1 = \left[\overrightarrow{NP}_{11}, \overrightarrow{NP}_{12}, \ldots \overrightarrow{NP}_{1n_1}\right]$, and $H_2 = \left[\overrightarrow{NP}_{21}, \overrightarrow{NP}_{22}, \ldots \overrightarrow{NP}_{2n_2}\right]$, where each $\overrightarrow{NP}$ is the phrase vector estimated by averaging the word vectors contained in the phrase and n_1 and n_2 are the respective lengths of the two lists. Next, we create $n_1 \times n_2$ phrase similarity matrix containing cosine similarities of phrase vectors contained in two phrase lists. Each row of this matrix represents a similarity distribution of a phrase in T_1 to all phrases in T_2. A phrase similarity vector of length n_2 is then constructed

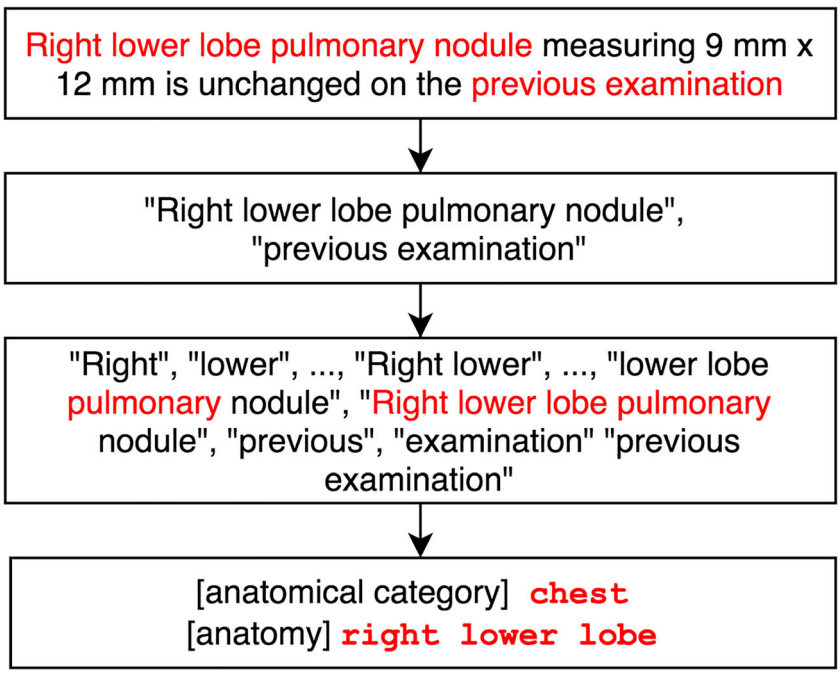

Fig. 3 Example of chunking and anatomy detection and categorization

Table 3 Confusion matrix of anatomy detection module

	Annotated with anatomical categories	Annotated without anatomical categories
Predicted with anatomical categories	735	66
Predicted without anatomical categories	15	351

by taking the 0.9 quantile of the distribution as the similarity of the phrase in each row: $PSV = [s_{1\cdot}, s_{2\cdot}, \ldots s_{n_2\cdot}]$. Finally, the similarity between the two texts $S(T_1, T_2) = \frac{1}{n_2} \sum_{i=1}^{n_2} s_i$ is computed by taking the average of *PSV*.

Classifier Features

Features used in Stage 1 are described below.

Comparison The comparison section of a radiology report often contains references to prior examinations related to the current report. When present, these references are strong indicators of an appropriate follow-up candidate, but they are not completely correlated. We define a binary feature that is set if the date of the source examination is present in the comparison section of the candidate report.

Modality Recommendation modality match is a binary feature that checks if the candidate modality matches any of the recommended modalities. In addition, a cross modality feature between the recommended modalities and the candidate modalities is also used because for certain anatomical regions a different modality than the one suggested in the source recommendation sentence may be an appropriate follow-up examination. For example, an MRI of the liver is usually an appropriate follow-up examination when a CT was previously recommended; this allows the model to extract inherent statistical relations between recommended modalities and other complementary modalities.

Patient Setting The patient setting is one of emergent, inpatient, or outpatient. In order to systematically capture the correlations between the two settings, the patient setting of the source examination and the candidate examination are combined into a categorical feature with the Cartesian product of the three possible labels.

Recommendation Interval Recommendations are characterized by an interval within which the follow-up examination should be ideally conducted. Examinations outside this interval are less likely to be appropriate follow-ups. We match candidates based on a function that provides a soft match of the recommendation interval:

$$\mathrm{I}(S_R, C_i) = \begin{cases} 1 & if \;\; R_{LB} \leq DATE\left(C_R^i\right) \leq R_{UB} \\ e^{-\alpha_l \left| \frac{DATE(C_i) - R_{LB}}{1 + R_{LB}} \right|^{\beta_l}} & if \;\; DATE(C_i) < R_{LB} \\ e^{-\alpha_u \left| \frac{DATE(C_i) - R_{UB}}{1 + R_{LB}} \right|^{\beta_u}} & if \;\; DATE(C_i) > R_{UB} \end{cases}$$

where, (α_l, β_l), and (α_u, β_u) are shape parameters for

```
IMPRESSION:

 Bilateral acute pulmonary emboli affecting
the bilateral upper lobes, middle lobe and
lingula. Emboli are involving subsegmental
arterial branches. Several filling defects
could be chronic in etiology.

 Filling defect suspicious for
acute/subacute pulmonary embolus involves
the right/upper lobe arterial branch.

 Bilateral subcentimeter pulmonary nodules
measure up to 7 mm on axial imaging, for
which follow-up CT is recommended in 3-6
months for a patient at high risk for
pulmonary malignancy.

 Bilateral hilar lymph nodes, measuring up
to 15 mm along short axis.

Small sliding hiatal hernia.
```

Source Report

```
IMPRESSION:

1. Multiple bilateral lung nodules are
stable. No new pulmonary nodules. Suggest
repeat chest CT in 6 months to assure
statbility.

2. Previously noted patchy groundglass and
ill-defined nodular opacities within the
right lung, likely representing resolved
infection.
```

Candidate Report

Fig. 4 Impression section of a source report and candidate report with similar contents

early and late examinations respectively. This function assigns a continuous matching score in (0, 1), where a value of 1 indicates that the candidate study is in the recommended interval, and a value of 0 to indicate outside interval. This function provides normalization of the time difference between the two studies based on the lower (R_{LB}) and upper bounds (R_{UB}), and a mechanism to set different decays for late and early follow-ups. In this work, these values were set to $\alpha_l = 0.5$, $\beta_l = 2$, $\alpha_u = 1$, *and* $\beta_u = 2$.

Section Matching We further define a class of features that uses the phrase-based text similarity method described earlier to compute relevance of the recommendation context and sections in the source report to the various sections of the candidate report.

A recommendation context is defined by tracing back maximally four sentences in a section until we find a phrase with a labeled anatomy. A four-sentence window was empirically found to be sufficient to isolate the anatomy if one existed. The phrases in this recommendation context are then matched to the findings, history, impression, and indication sections of the candidate report individually and to the entire candidate report.

In addition, we also use similarity between the source and candidate reports. In addition to similarity between the entire source and candidate reports, we estimate individual similarities between the sections of the respective reports represented here by (source report section, candidate report section): (findings, findings), (history, history), (impression, impression), (impression, history), and (impression, indication).

Anatomy Matching While the section matching features help in identifying similar reports, in order to achieve precision in identifying follow-up candidates, we use features that match anatomy labels identified using the anatomy detection module. Specifically, we match the anatomy labels in the recommendation context to the anatomy labels found in the impression, indication and procedure sections of the candidate report.

Results

We evaluated the performance of the algorithm on the reconciled dataset described previously. This dataset contains 559 source exams (SA) that contain recommendations. Out of these, the radiologists annotated 387 SAs with a subsequent follow-up candidate exam (CA). The remaining 172 SAs do not have any follow-up accessions and hence these recommendations were not followed-up. The algorithm's task is to classify each of the 559 source accessions (SAs) as followed-up or not followed-up and to identify the correct CA for a SA that is followed-up.

To analyze the performance of the model we categorize SAs as follows into mutually exclusive groups:

1. Correct follow-up (CFU): the model predicts the correct CA for a followed-up SA
2. Wrong follow-up (WFU): the model predicts the incorrect CA for a followed-up SA
3. Missed follow-up (MFU): the model identifies no CA for a followed-up SA
4. Correct no follow-up ($C\overline{FU}$): the model correctly identifies no CA for a not followed-up SA
5. False follow-up (FFU): the model incorrectly identifies a CA for a not followed-up SA

In order to compute precision and recall for the classifier, we note that CFU implies a true positive label, $C\overline{FU}$ is true negative, MFU is false negative, and WFU combined with FFU is false positive. The precision (P_{FU}) and recall (R_{FU}) for followed-up SAs and the precision $\left(P_{\overline{FU}}\right)$ and recall $\left(R_{\overline{FU}}\right)$ for not followed-up SAs are then used to compute the F1-Score and accuracy.

As a reference for evaluation of the machine learning model, we note the pair-wise annotator agreements (F1-score) before reconciliation were as follows: annotator (Ann) 1–Ann 2: 0.861, Ann 2–Ann 3: 0.868, Ann 3–Ann 1: 0.853.

As a baseline for comparison, we consider a model that selects the earliest candidate that matches only the recommended modality. For the evaluation of model performance, we split the 559 SAs into 70% that are used for training and the remaining 30% for testing. The aggregate results for the baseline model, the proposed ERT model and the analysis differences for two of the human annotators aggregated over 50 such independent splits are shown in Table 4.

As seen in Table 4, selecting the earliest candidate matching the recommended modality provides high recall but suffers from low precision. The ERT classifier improves on both the precision of the baseline and compares favorably to the inter-annotator performance in overall performance.

The features described above previously were encoded into a total of 60 bases for use in the classifier. The top 20 bases are shown in Fig. 5, which represents the average feature importance over the 50 splits.

As seen in the figure, matches on recommended modality and interval are the top two predictors of correctly identifying the follow-up exam. The next important feature is the match of the anatomy in the recommendation context to the content in the impression, findings, and procedure sections of the

Table 4 Comparison of inter-annotator agreement vs. the ERT machine learning model

	Accuracy	FAVG	$F_{\overline{FU}}$	F_{FU}	$P_{\overline{FU}}$	$R_{\overline{FU}}$	P_{FU}	R_{FU}
Ann1–Ann3	0.812	0.795	0.736	0.853	0.974	0.593	0.752	0.987
ERT model	0.765	0.753	0.699	0.807	0.806	0.617	0.744	0.882
Baseline	0.581	0.551	0.436	0.666	0.778	0.304	0.529	0.900

candidate examination (in that order). The similarity between the recommendation context to the impression section and the overall similarity between the two reports are also strong predictors of appropriate follow-up.

Discussion

The follow-up matching algorithm provides a means for radiologists and radiology department administrators to determine if patients have completed the clinically appropriate follow-up imaging study. A system to proactively send reminders to referrers, primary care providers, and patients can ensure that patients receive the timely care they need.

In this paper, we have presented a methodology that can be used to determine which exam, if any, is the most likely clinically appropriate follow-up exam for a given follow-up imaging recommendation. Using a multi-year clinical dataset from a tertiary health system, we have presented a scalable approach that can determine the follow-up exam with reasonable accuracy using:

1) Study meta-data (e.g., patient setting, modality),
2) Recommendation context in source exam (e.g., recommended anatomy, modality, and interval),
3) Report text-based similarity features (e.g., similarity between the recommendation sentence and impression section and/or the reason for exam of the candidate report), and
4) Features such as time difference between the source and the candidate exams.

Few studies have focused on scalable methods to identify timely follow-up studies. Al-Mutairi et al. reported that language in the report suggestive of doubt did not affect the timeliness of the follow-up [19]. Wandtke et al. reported on steps undertaken that improved the rate of follow-ups [20]. For instance, by asking radiologists to dictate certain phrases into the reports, tracking systems can automatically extract follow-up recommendations that contain these phrases with an explicit follow-up interval and create alerts when a follow-up examination is due [20]. Similarly, an explicit score can be assigned to indicate the degree of suspicion for lesion malignancy and the need for follow-up [7]. However, to the best of the authors' knowledge, this is the first effort to automatically identify the most relevant follow-up study for follow-up imaging recommendations from free text reports

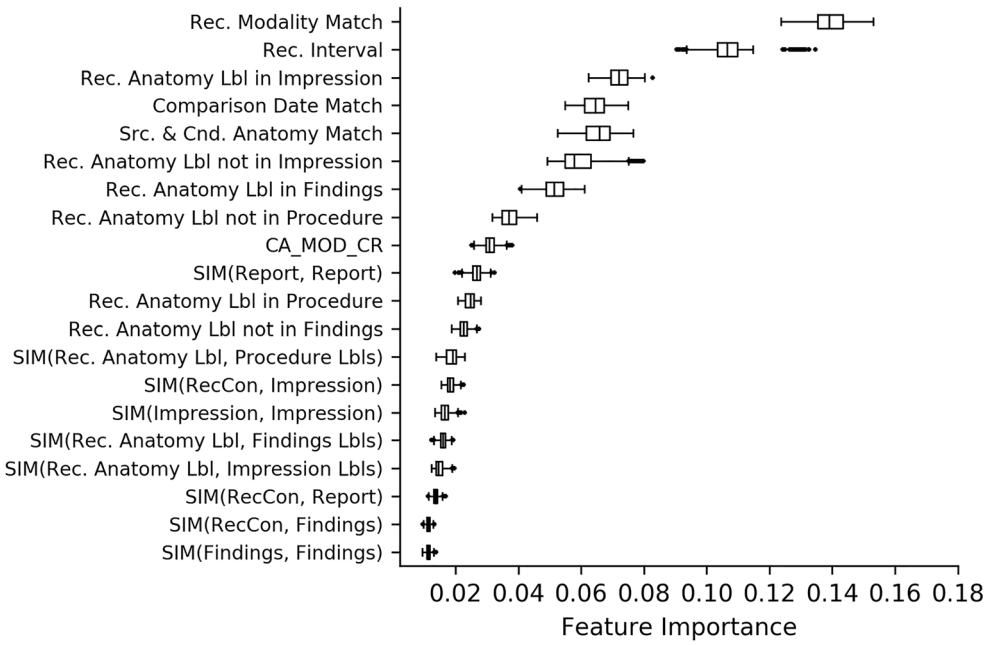

Fig. 5 Feature importance

in routinely produced radiology reports. In a previous paper, we describe an application of the proposed algorithm to determine adherence of imaging recommendations using a large production data set of 2,972,164 exams that covers 7 years of study [21].

The presented model has an F-score of 0.807 for followed-up SAs. This was deemed acceptable given that the corresponding inter-radiologist F-score ranged from 0.853 to 0.868%.

The relative importance of the top features for the classifier confirms intuitive expectations that the follow-up examination should match the recommended modality and be performed within the recommendation time interval, but as expected, the modality match is not exact (e.g., a CT of the chest would be considered to be appropriate follow-up for a pulmonary finding when a chest radiograph was previously recommended). The importance of anatomy matching in the impression section can be explained by the fact that radiologists do elaborate on pertinent anatomy from the recommendation context of the source examination in the impression section of a report.

Although the algorithm accuracy is reasonable and provides a scalable approach to determine if a recommendation has been followed-up or not, there are limitations to the method described herein.

First, recommendations for follow-up imaging are couched in language that limits the ability to accurately judge the clinical necessity of follow-up (e.g., "If clinically indicated, follow-up CT could be performed in 4–6 weeks to document resolution") [22]. Follow-up recommendations with explicitly specified imaging modalities and time-intervals may imply the need for a follow-up, although we did not investigate if such recommendations lead to a higher follow-up rate in the current data.

Second, the algorithm relies on having access to longitudinal radiology reports to determine if follow-up was performed. In hospital settings such as tertiary care hospitals, there are often referrals for specialized procedures from areas outside of the primary catchment area for the radiology department, including out-of-state referrals, and it is possible that such exams may get flagged as having no follow-up by the algorithm although the patients may have followed up on the imaging recommendation with a more local radiology facility.

Third, there are some inherent limitations due to the variability of natural language and clinical complexity—three representative examples based on a manual review of results from one of the fifty splits are:

1) In one of the wrong follow-up cases, the predicted follow-up study and the true follow-up study both received high scores (0.96 and 0.83 respectively), both matched the recommended modality, had comparison dates to match the source study, and described the primary findings of pulmonary nodules in each case. However, the true follow-up marked by the radiologist was performed earlier than the recommended time-frame that justified its lower matching score of 0.83 compared with the predicted study with 0.96.
2) In one of the missed follow-up cases, the true follow-up study marked by the radiologists was performed 1 year later than the recommended follow-up interval and had an incorrect modality as confirmed by the radiologist. The algorithm match score for this follow-up study was 0.49. Similarly, in another Missed Follow-up case, the true follow-up study was within the recommended time-frame but only received a score of 0.39. The true follow-up had a comparison section, matched the recommended modality but did not have any text in the impression section to derive any meaningful text similarity features;
3) In one of the false follow-up cases, the radiologists correctly did not pick the predicted study as a true follow-up because the clinically recommended exam should have been a "with contrast" study, and the classifier did not use that as a feature. On the other hand, one of the false follow-up cases was reclassified as a true follow-up—this was missed by the radiologists as being the follow-up candidate as the report text contained the text for a CT liver study followed by text for a CT chest (which was the true follow-up).

We have demonstrated the value of automated follow-up matching using radiology reports; however, we believe that the proposed methodology of using machine learning and natural language processing-based processing can be extended to other domains as well, for instance, to determine if a clinical follow-up has been completed as recommended after a benign breast biopsy since these patients are at elevated risk for the subsequent development of breast cancer.

Conclusion

- A machine learning and natural language processing-based algorithm can be developed to automatically identify clinically appropriate follow-up studies with reasonable accuracy.
- The precision and recall performance parameters show promise that such an algorithm can be integrated into a routine follow-up tracking system for radiology departments to support quality improvement initiatives.

Compliance with Ethical Standards

Conflicts of Interest Authors TM, VH, SD, and CH are employees of Philips working in collaboration with the University of Washington, Department of Radiology under an industry-supported master research

agreement. This manuscript details original research performed under this agreement in compliance with the Sunshine Act, but does not employ an existing Philips product. Authors TM and CH also have Adjunct Faculty Appointments with the University of Washington.

References

1. Sistrom CL, Dreyer KJ, Dang PP, Weilburg JB, Boland GW, Rosenthal DI, Thrall JH: Recommendations for additional imaging in radiology reports: multifactorial analysis of 5.9 million examinations. Radiology 253(2):453–461, 2009
2. Larson PA, Berland LL, Griffith B, et al. Actionable findings and the role of IT support: report of the ACR actionable reporting work group
3. Kadom N, Doherty G, Close M et al.: Safety-net academic hospital experience in following up noncritical yet potentially significant radiologist recommendations. Am J Roentgenol 209(5):982–986, 2017
4. Blagev DP, Lloyd JF, Conner K, Dickerson J, Adams D, Stevens SM, Woller SC, Evans RS, Elliott CG: Follow-up of incidental pulmonary nodules and the radiology report. J Am Coll Radiol 11(4):378–383, 2014
5. Kulon M, "Lost to Follow-Up: automated Detection of Patients Who Missed Follow-Ups Which Were Recommended on Radiology Reports", SIIM Conference Proceedings 2016.
6. Sloan CE, Chadalavada SC, Cook TS et al.: Assessment of follow-up completeness and notification preferences for imaging findings of possible cancer: what happens after radiologists submit their reports? Acad Radiol 21(12):1579–1586, 2014
7. Cook TS, Lalevic D, Sloan C, Chadalavada SC, Langlotz CP, Schnall MD, Zafar HM: Implementation of an Automated Radiology Recommendation-Tracking Engine for Abdominal Imaging Findings of Possible cancer. J Am Coll Radiol 14(5): 629–636, 2017
8. Dang PA, Kalra MK, Blake MA, Schultz TJ, Halpern EF, Dreyer KJ: Extraction of Recommendation Features in Radiology with Natural Language Processing: Exploratory Study. AJR 191:313–320, 2008
9. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH: Automatic identification of critical follow-up recommendation sentences in radiology reports. AMIA Annual Symp Proc 2011:1593–1602
10. Dutta S, Long WJ, Brown DF, Reisner AT: Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. Ann Emerg Med 62(2):162–169, 2013
11. Oliveira L, Tellis R, Qian Y et al., Follow-up Recommendation Detection on Radiology, Reports with Incidental Pulmonary Nodules. Stud Health Technol Inform. 2015; 216:1028.
12. Gunn M.L., Yetisgen M, Lehnert B.E. et al., Automating Radiology Quality and Efficiency Measures with Natural Language Processing, Radiological Society of North America 2015 Scientific Assembly and Annual Meeting.
13. Gunn M.L., Lehnert B.E., Hall C et al., Impact of Patient Location and Radiology Subspecialty on Imaging Follow-up Recommendation Rate, Radiological Society of North America 2015 Scientific Assembly and Annual Meeting.
14. Mabotuwana T, Hall C, Dalal S, Tieder J, Gunn M. Extracting follow-up recommendations and associated anatomy from Radiology Reports. in 16th World Congress on Medical and Health Informatics (MedInfo2017). Aug 2017. Hangzhou.
15. Geurts P, Ernst D, Wehenkel L: Extremely randomized trees. Machine Learning 63(1):3–42, 2006
16. Rosse C, Mejino JLV: A reference ontology for biomedical informatics: the Foundational Model of Anatomy. Journal of Biomedical Informatics 36(6):478–500, 2003
17. Mikolov T, Sutskever I, Chen K et al., (2013) Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26 (NIPS 2013), p. 3111–3119
18. Blacoe W, Lapata M, A Comparison of Vector-based Representations for Semantic Composition, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, p. 546–556.
19. Al-Mutairi A, Meyer AND, Chang P et al.: Lack of Timely Follow-Up of Abnormal Imaging Results and Radiologists' Recommendations. J Am Coll Radiol 12(4):385–389, 2015
20. Wandtke B, Gallagher S: "Closing the Loop: A Radiology Follow-up Recommendation Tracking System," RSNA Quality Storyboards 2016.
21. Mabotuwana T, Hombal V, Dalal S, Hall C, Gunn M. Determining Adherence to Follow-up Imaging Recommendations. Journal of American College of Radiology 2018 Mar; p. 422–428
22. Gunn ML, Lehnert BE, Hall CS, et al. Use of conditional statements in radiology follow-recommendation sentences: relationship to follow up compliance. Radiological Society of North America 101st Scientific Assembly and Annual Meeting; Chicago. 2015.