Annotation of Clinically Important Follow-up Recommendations in Radiology Reports

Meliha Yetisgen[1,2], Prescott Klassen[2], Lucas H. McCarthy[3], Elena Pellicer[4], Thomas H. Payne[4,5], Martin L. Gunn[6]
Department of Biomedical Informatics and Medical Education[1], Department of Linguistics[2], Department of Neurology[3], School of Medicine[4], Information Technology Services[5], Department of Radiology[6]
University of Washington
Seattle, WA
melihay,klassp,lucasmc,pellicer,tpayne,marting@uw.edu

Abstract

Communication of follow-up recommendations when abnormalities are identified on imaging studies is prone to error. The absence of an automated system to identify and track radiology recommendations is an important barrier to ensuring timely follow-up of patients especially with non-acute incidental findings on imaging studies. We are in the process of building a natural language processing (NLP) system to identify follow-up recommendations in free-text radiology reports. In this paper, we describe our efforts in creating a multi-institutional radiology report corpus annotated for follow-up recommendation information. The annotated corpus will be used to train and test the NLP system.

1 Introduction

A radiology report is the principal means by which radiologists communicate the findings of an examination to the referring physician and sometimes the patient. With the dramatic rise in utilization of medical imaging in the past two decades, health providers are challenged by the optimal use of clinical information while not being overwhelmed by it. Based on potentially important observations the radiologist may recommend specific imaging tests or a clinical follow-up in the narrative radiology report. These recommendations are made for several potential reasons. The radiologist may recommend further investigation to clarify the diagnosis or exclude potentially serious, but clinically expected disease. Secondly, the radiologist may unexpectedly encounter signs of potentially serious disease on the imaging study that they believe require further investigation. Thirdly, the radiologist may recommend surveillance of disease to ensure an indolent course. Finally, a radiologist may provide advice to the referring physician about the most effective future test(s) specific to the patient's disease or risk factors.

The reliance on human communication, documentation, and manual follow-up is a critical barrier to ensuring that appropriate imaging or clinical follow-up occurs. The World Alliance of Patient Safety, a part of the World Health Organization, recently identified poor test results follow-up as one of the major processes contributing to unsafe patient care[1].

There are many potential failure points when communicating and following up on important radiologic findings and recommendations: (1) Critical findings and follow-up recommendations not explicitly highlighted by radiologists: Although radiologists describe important incidental observations in reports, they may or may not phone an ordering physician. If these recommendations "fall through the cracks" patients may present months later with advanced disease (e.g., metastatic cancer). (2) Patient mobility: When patients move between services in healthcare facilities, there is increased risk during "hand-offs" of problems with test result follow-up and continuity of care (Callen et al., 2011). (3) Heavy workload of providers: Physicians and other pro-

[1] World Alliance for Patient Safety. Summary of the Evidence on Patient Safety: Implications for Research. Geneva: World Health Organization, 2008. Accessed: 3.13.2015. Available at: http://gawande.com/documents/WHOGuidelinesforSafeSurgery.pdf

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 50–54,
Lisbon, Portugal, 17 September 2015.

viders have to deal with a deluge of test results. A survey of 262 physicians at 15 internal medicine practices found that physicians spend on average 74 minutes per clinical day managing test results, and 83% of physicians reported at least one delay in reviewing test results in the previous two months (Holden et al., 2004). However, it is vital that these results, particularly if they are unexpected, are not lost to follow-up. In patients who have an unexpected finding on a chest radiograph, approximately 16% will eventually be diagnosed with a malignant neoplasm (Poon et al., 2004).

These examples indicate an opportunity to develop a systematic approach to augmenting existing channels of clinical information for preventing delays in diagnosis. The goals of our research are to: (1) define clinically important recommendations in the context of radiology reports and (2) create a large-scale radiology report corpus annotated with recommendation information. The corpus will be used to build an automated system that will extract recommendation information so that reports can be flagged visually and electronically.

2 Related Work

Identifying follow-up recommendation information in radiology reports has been previously studied by other researchers. Dreyer et al. processed 1059 radiology reports with Lexicon Mediated Entropy Reduction (LEXIMER) to identify the reports that include clinically important findings and recommendations for subsequent action (Dreyer et al., 2005). The same research group performed a similar analysis on a database of radiology reports covering the years 1995-2004 (Dang et al., 2008). From that database, they randomly selected 120 reports with and without recommendations. Two radiologists independently classified those selected reports according to the presence of recommendation, time-frame, and imaging-technique suggested for follow-up examination. These reports were analyzed by an NLP system first for classification into two categories: reports with recommendations and reports without recommendations. The reports with recommendations were then classified into those with imaging recommendations and those with non-imaging recommendations. The recommended time frames were identified and normalized into a number of days. The authors reported 100% accuracy in identifying reports with and without recommendations. In 88 reports with recommendation, they reported 0.945 precision in identifying temporal phrases, and 0.932 in identifying recommended imaging tests. In a follow-up study, the authors analyzed the rate of recommendations by performing a statistical analysis on 5.9 million examinations (Sistrom et al., 2009). In all three papers, they reported high overall performance values; however, the authors presented their text processing approach as a black box without providing nec-

01 CT ABDOMEN AND PELVIS WITH INTRAVENOUS CONTRAST
02 HISTORY:
03 Prostate CA-Prostate CA Surveillance
04 COMPARISON: None
05 CONTRAST: iv contrast was used. Positive oral contrast was administrated
06 TECHNIQUE:
07 Region of interest: Abdomen-Pelvis
08 Superior Extent: Diaphragm. Inferior Extent: Symphysis Pubis
09 ……..
10 FINDINGS:
11 Lung bases: A 6-mm nodule is noted in the peripheral left lung base (image 9, series 2). There is a focal area of
12 atelectasis in the anterior right lung base.
13 Pleura: No pleural effusions or thickening.
14 Included heart: No gross abnormality..
15 Liver: Normal
16 Portal veins: Normal.
17 Gallbladder and bile ducts: Normal
18 Spleen: Normal
19 Aorta and IVC: There is atherosclerotic calcification of the aorta.
20 There is a small focus of ulcerated plaque in the infrarenal aorta (image 49, series 2).
21 Stomach, duodenum and small bowel: Normal
22 ……..
23 IMPRESSION:
24 1. Incidental 6-mm left lung nodule. Follow-up chest CT is recommended in 6 months.
25 2. A few prostatic calcifications are noted. No CT evidence of metastatic prostate cancer.
26 3. Small ulcerated atheromatous plaque in the infrarenal aorta.

Figure 1: Example radiology report with follow-up recommendation

essary information required to replicate their methods.

3 Follow-up Recommendations in Radiology Reports

In this research, we define a *follow-up recommendation* as a statement made by the radiologist in a given radiology report to advise the referring clinician to further evaluate an imaging finding by either other tests or further imaging. Figure 1 presents a radiology report with such a follow-up recommendation (Line 24: Incidental 6-mm left lung nodule. *Follow-up chest CT is recommended in 6 months*).

Under the supervision of a radiologist and an internal medicine specialist, we analyzed a small set of radiology reports with different modalities and grouped the follow-up recommendations under the following four non-overlapping categories.

Category 1: Non-contingent clinically important recommendation: An advisory statement that could result in mortality or significant morbidity if appropriate clinical assessment, diagnostic or therapeutic follow-up steps are not followed.
Case example: Incidental lung mass suspicious for malignancy on a trauma CT of the abdomen.
Follow-up recommendation example: *CT chest is recommended to further evaluate the lung mass*.

Category 2: Contingent clinically important recommendation: Similar to (a), but the statement is conditional on the presence of a clinical condition.
Case example: Adrenal mass identified on a CT of the abdomen and pelvis for appendicitis.
Follow-up recommendation example: *If the patient has a history of malignancy, consider biochemical testing and an adrenal mass protocol CT for further evaluation.*

Category 3: Clinically important recommendation likely reported: Similar to (a) and (b), but considered to be unlikely not to be reported in communication between radiologist and clinician.
Case example: A distal radius fracture was identified on a previous week's x-ray of patient's hand. A follow-up x-ray of the hand is requested to rule out possible additional scaphoid fracture.
Follow-up recommendation example: *L distal radius fracture x 1 week, please also follow-up to rule out scaphoid fracture compared with last week's x-rays*.

Category 4: Clinically unimportant recommendation: An advisory statement that is unlikely to result in mortality or significant morbidity if appropriate clinical assessment, diagnostic or therapeutic follow-up steps are not followed, and/or a low probability that the recommendation would be overlooked.
Case example: Following trauma, a radiograph demonstrates a probable non-displaced fracture of the mid ulna.
Follow-up recommendation example: *Consider an MRI of the forearm if diagnostic certainty is desired.*

To capture the main attributes of follow-up recommendations, we created a simple template with three entities; reason for recommendation (e.g., *incidental 6-mm left lung nodule*), recommended test (e.g., *chest CT*), and time-frame (e.g., in *6 months*). We use the follow-up recommendation categories and template to annotate a large scale radiology corpus that will be explained in the following sections.

4 Corpora for Follow-up Recommendations

4.1 Pilot Corpus

Dataset: In previous work, we created a corpus of radiology reports composed of 800 de-identified radiology reports extracted from the radiology information system of our institution (Yetisgen-Yildiz et al., 2013). The reports represented a mixture of imaging modalities, including radiography computer tomography (CT), ultrasound, and magnetic resonance imaging (MRI). The distribution of the reports across imaging modalities is listed in Table 1.

Imaging modality	Frequency
Computer tomography	486
Radiograph	259
Magnetic resonance imaging	45
Ultrasound	10
Total	800

Table 1: Distribution of reports in pilot corpus.

Annotation Guidelines: We annotated this dataset prior to defining different categories of follow-up recommendations. In this annotation task, we asked the annotators simply to highlight the boundaries of sentences that include any follow-up recommendation.

Annotation Process: Two annotators, one radiologist and one internal medicine specialist, went through each of the 800 reports and marked the sentences that contained follow-up recommendations. Out of 18,748 sentences in 800 reports, the radiologist annotated 118 sentences and the clinician annotated 114 sentences as recommendation. They agreed on 113 of the sentences annotated as recommendation. The inter-rater agreement was 0.974 F-score.

4.2 Multi-institutional Radiology Corpus

We extended our pilot dataset of 800 reports with a much larger set of 745,058 radiology reports from three different institutions including University of Washington Medical Center, Harborview Medical Center, and Seattle Cancer Care Alliance. The corpus covers the full range of imaging modalities, including radiographs, computed tomography, ultrasound, and magnetic resonance imaging (Table 2).

Imaging modality	Frequency
Computed Radiography	413,889
Computed Tomography	146,181
Digital Fluoroscopy	12
Digital Radiography	1,626
Magnetic Resonance Imaging	52,127
Nuclear Medicine	12,895
Portable Radiography	6,166
Portable Radiography	4,121
Fluoroscopy	27,239
Ultrasound	68,999
Angio-Interventional	11,803
Total	745,058

Table 2: Distribution of reports in multi-institutional radiology corpus.

We excluded the Mammography modality, which was comprised of 37,754 reports because a specific follow-up and alert system was already in place.

Annotation Guidelines: We designed the annotation task to operate on two levels; sentence level and entity level. At the sentence level, the annotators mark the boundaries of recommendation sentences and label each marked sentence with one of the four recommendation categories: (1) *non-contingent clinically important recommendation*, (2) *contingent clinically important recommendation*, (3) *clinically important recommendation likely reported*, and (4) *clinically unimportant recommendation*. At the entity level, the annotators mark the three attributes of recommendation information presented in the marked sentences: (1) *reason for follow-up recommendation*, (2) *recommended follow-up test*, and (3) *time-frame for follow-up test*.

Annotation Process: Because manual annotation is a time-consuming and labor-intensive process, we could annotate only a small portion of our large radiology corpus. The percentage of reports that include recommendation sentences is quite low—about 15% at our institution. To increase the number of reports with recommendations in the annotated set, rather than randomly sampling, we built a high recall (0.90), low precision (0.35) classifier trained on the pilot dataset described in section 4.1. The details of this baseline classifier can be found in our prior publication (Yetisgen-Yildiz et al., 2013). We ran our baseline classifier on un-annotated reports and only sampled reports for manual annotation from the reports our classifier identified as positive for follow-up recommendations. Because the classifier was high recall but low precision, it identified many false positives. The filtering of reports using a classifier reduced the number of reports our human annotators needed to review, expediting the annotation process.

At the sentence level, one radiologist and one neurologist review the classifier-selected reports with system generated follow-up recommendation sentences highlighted. The annotators correct the system generated sentences and/or highlight new sentences if needed. They associate each highlighted sentence with one of the four types described in Section 3.

At the entity level, one neurologist and one medical school student annotate the entities (reason for recommendation, recommended test, and time frame) in reports annotated in a previous stage at the sentence level with follow-up recommendations.

Inter-annotator Agreement Levels: At the sentence level, we measured the inter-annotator agreement on a set of 50 reports featuring at least one system-generated recommendation identified by our high recall classifier from a randomly selected collection of one thousand reports. Our annotation process required annotators to re-label all sentences that were initially identified by the system as a recommendation with the four type-specific labels described in Section 3. They could label the sentence as *Incorrect* if they believed the system had wrongly identified a recommendation sentence and they could also label a new

recommendation sentence if they believed it had not been identified correctly by the system. The inter-rater agreement levels were kappa 0.43 and 0.59 F1 score. To resolve the disagreements, we scheduled multiple meetings. One of our observations during those meetings was that none of the new recommendation sentences introduced by either annotator were identified by the other. In our review, both annotators agreed that the majority of the new recommendations the other introduced were correct. We adjusted our annotation guidelines to add rules to help decide if and when a new sentence should be identified as a recommendation.

At the entity level, agreement levels were 0.78 F1 for reason, 0.88 F1 for test, and 0.84 F1 for time frame.

Our annotation process is on-going. The annotators completed the annotation of 567 radiology reports using updated guidelines based on the inter-annotator agreement stage. They highlighted 265 sentences as category 1, 90 sentences as category 2, 222 sentences as category 3, and 160 sentences as category 4. At the entity level, for 225 recommendation sentences, the annotators highlighted 207 text spans as reason, 314 text spans as test, and 71 text spans as time-frame.

5 Conclusion

In this paper, we described our efforts in creating a large scale radiology corpus annotated for follow-up recommendations. We are in the process of building a text processing system based on our current annotated corpus.

References

Callen J, Georgiou A, Li J, Westbrook JI. The safety implications of missed test results for hospitalized patients: a systematic review. BMJ Qual Saf. 2011;20(2):194-9.

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. 2008; 191:313-20.

Dreyer KJ, Kalra MK, Maher MM, Hurier AM, Asfaw BA, Schultz T, Halpern EF, Thrall JH. Application of Recently Developed Computer Algorithm for Automatic Classification of Unstructured Radiology Reports: Validation Study. Radiology. 2005; 234:323-39.

Holden WE, Lewinsohn DM, Osborne ML, Griffin C, Spencer A, Duncan C, Deffebach ME. Use of a clinical pathway to manage unsuspected radiographic findings. Chest. 2004;125(5):1753-60.

Poon EG, Gandhi TK, Sequist TD, Murff HJ, Karson AS, Bates DW. "I wish I had seen this test result earlier!": Dissatisfaction with test result management systems in primary care. Arch Intern Med. 2004;164(20):2223-8.

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. 2009; 253(2):453-61.

Xia F, Yetisgen-Yildiz M. Clinical Corpus Annotation: Challenges and Strategies. Proceedings of Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) of the International Conference on Language Resources and Evaluation (LREC), Istanbul, May, 2012.

Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. A Text Processing Pipeline to Extract Recommendations from Radiology Reports. J Biomed Inform., 2013;46(2):354-362.