Automated Evaluation of Radiology Reports: Defining and measuring metrics to standardize quality assurance

School of Medicine and Public Health
UNIVERSITY OF WISCONSIN–MADISON

John Renfrew*, Advisor: Dr. Jason Stephenson, M.D.**

*University of Wisconsin – School of Medicine and Public Health

**UW – SMPH, Department of Radiology, Assistant Professor of Radiology

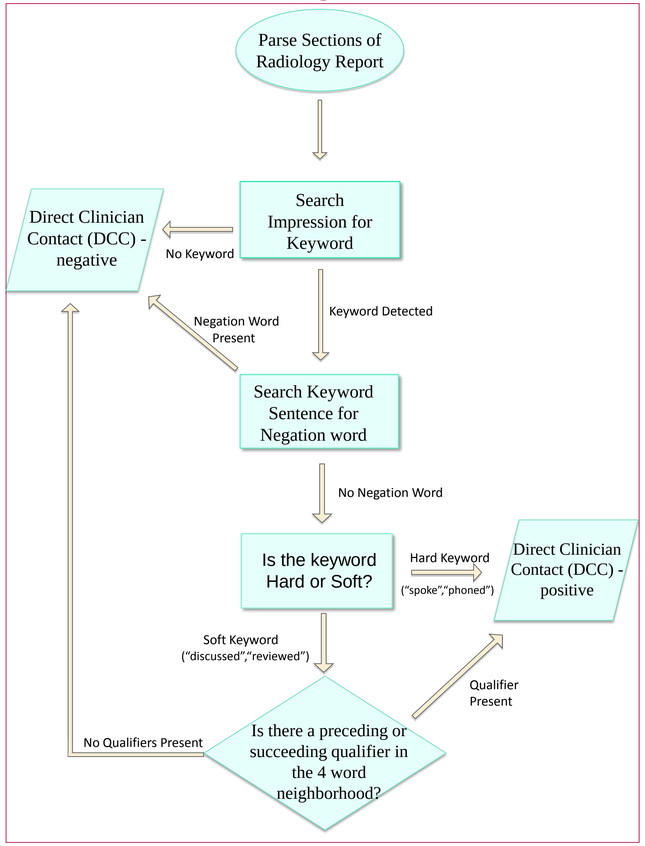
BACKGROUND

Medical staff credentialing committees and accountable care organizations (ACOs) are requiring more and more radiologist performance measurements (metrics). Using human graders to evaluate large numbers of radiology reports to create performance metrics is costly. We devised an automated process for generating performance metrics and evaluated the automated process using human grading as the reference standard.
Particularly, we choose to investigate aspects of a radiological report that suggest increased communication between health care workers. The best measure of this communication includes non-routine contact between the radiologist and ordering provider (e.g. phone call). We are particularly interesting in the occurrence of non-routine contact in the setting of an additional imaging recommendation by the reporting radiologist.

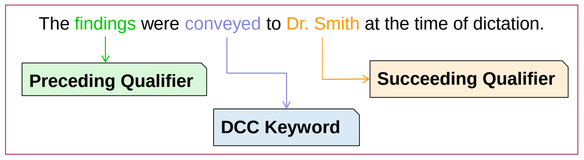
MATERIALS & METHODS

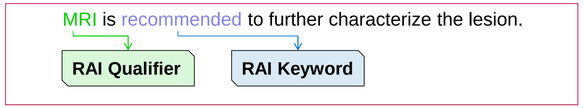
We developed an automated process (computer algorithm) for scanning radiology reports for two features: recommendation for additional imaging (RAI) and documentation of clinician contact (DCC). We then compared the automated process with human grading of the same reports. The report database consisted of 76,814 reports generated by an independent, private, general practice radiology group.
We used the Java programming language to construct a computer algorithm to automatically process radiology reports. The first step in this process was to use 1,000 reports to develop a list of keywords to create the RAI and DCC algorithms. We used an iterative process of comparing the results of the computer algorithm to human grading in novel batches of 250 reports, warranting a total of 5,500 reports. We refined the algorithm to improve performance via four methods:
1. Dividing DCC keywords into two categories: *hard* keywords (e.g., “spoke” and “phoned”) which automatically qualified the report as positive, and *soft* keywords (e.g., “discussed” and “reviewed”) which required the presence of “Qualifier” words. All RAI keywords were considered *soft*.
2. Definition of the Qualifier words: for DCC, we created separate lists of words appearing before and after (in a four word neighborhood) the keyword. For RAI, we created a single list but used a larger neighborhood.
3. Limiting the RAI search to the *impression* field of the reports.
4. Employing sentence level analysis to enhance the accuracy of the algorithm.

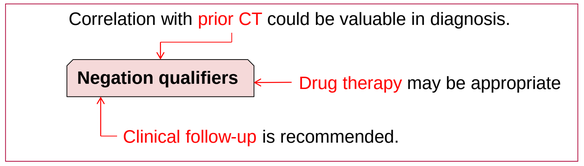
DCC Algorithm



DCC & RAI EXAMPLES







RESULTS

Recommended additional imaging (RAI)	
Keywords	Follow-up, indicated, helpful, necessary, consider, needed, valuable, provide useful, performed, obtained, considered, recommendation, recommend, better characterization, repeat, further radiographic, add more information, examination may allow, may be appropriate, would be needed, evaluation, suggest, correlation with
Qualifying words	Any imaging modality including: Computed tomography, CT, magnetic resonance imaging, MRI, ultrasound, US, cerebral angiogram, CA, film, radiograph, plain film, positron emission tomography, PET, bone scan, scan, scanning
Negation list	Previous, clinical follow-up, clinical correlation, correlation with patient’s, physical exam, injection, correlation with clinical, follow-up clinically, clinically is recommended, on a clinical basis, clinical course, drug therapy, colonoscopy, correlation with site, recent, old films, cytology, prior study, prior studies, no further screening, no further imaging, no further study, no additional testing, no additional screening, no additional imaging, no additional study, no additional testing, no screening, no imaging, no study, no testing
Documented Clinician Contact (DCC)	
Keywords	Spoke, talked, discussed, reviewed, called, phoned, notified, conveyed, communicated, reported, relayed, made (certain situations)
Preceding qualifying words	I, radiologist, doctor, ordering provider, call, was, findings, results
Succeeding qualifying words	With, to, given, Dr., findings, by
Negation list	Rad tech, technologist, patient, sonographer

RESULTS – Recommended Additional Imaging (RAI)

	Test Negative	Test Positive
Reference Standard Negative	2323	8
Reference Standard Positive	13	122

Table 2: Statistical summary of RAI testing
Accuracy = 99.1% (95% CI: 98.6% - 99.4%);Sensitivity (recall) = 90.3% (95% CI: 84.1% - 94.3%);Specificity = 99.6% (95% CI: 99.3% - 99.8%);Positive predictive value (precision) = 93.8% (95% CI: 88.1% - 97.0%);Negative predictive value = 99.4% (95% CI: 99.0% - 99.7%); F1 Score: 0.921

RESULTS – Direct Clinician Contact (DCC)

	Test Negative	Test Positive
Reference Standard Negative	2258	2
Reference Standard Positive	3	203

Table 3: Statistical summary of DCC testing
Accuracy = 99.8% (95% CI: 99.5% - 99.9%);Sensitivity (recall) = 98.5% (95% CI: 95.6% - 99.7%);Specificity = 99.9% (95% CI: 99.6% - 99.9%);Positive predictive value (precision) = 99.0% (95% CI: 96.2% - 99.4%);Negative predictive value = 99.8% (95% CI: 99.5% - 99.9%);F1 score: 0.988

DISCUSSION

Our results indicate that an automated process for detecting RAI and DCC can demonstrate good when using human grading as reference standard. Using an iterative process, we achieved greater than 99% accuracy in report grading. The F1 scores (a summary measure combining positive predictive value and sensitivity) were .921 for RAI and 0.988 for DCC.
The ability to accurately measure RAI and DCC across a wide range of radiologists reporting on multiple modalities in an automated, cost-effective measure makes it possible to produce individual and group measurements of radiologist performance (metrics).
The implications of successful automated report grading include the ability to provide individual and group metrics to client hospital and accountable care organizations, and to identify differences in radiologist performance to direct efforts at practice improvement. In this regard, RAI and DCC may become analogous to performance metrics in mammography (e.g., recall rate, biopsy rate, and cancer detection rate) that have recognized national benchmarks. It is also possible to use a slight modification of the process reported here to create an additional metric, namely, the percentage of cases in which RAI is also DCC: that is, the percentage of the time that the radiologist discusses with the referring clinician the recommendation for further imaging. This contact is critically important, providing a safety net, and ensuring the follow-up recommendation does not go unrecognized.
In conclusion, we developed an automated process for the detection of recommended additional imaging (RAI) and direct clinician contact (DCC) in a large general practice radiology group that was accurate when compared to human grading. Such an automated process allows for calculation of radiology performance metrics in a cost-effective manner, which will be in greater and greater demand by medical staff credentialing processes and accountable care organizations over the next several years. Future work involves the addition of features to the algorithm in an attempt to make it an even better Quality Assurance tool. Namely, itemized reporting and acknowledgement of the direct indication for the report.

REFERENCES

Allen B, Levin DC, Brant-Zawadzki M et al. ACR white paper: strategies for radiologists in the era of health care reform and accountable care organizations: a report from the ACR future trends committee. J Am Coll Radiol 2011;8:309-317.
Hripcsak G, Friedman C, Alderson PO et al. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995;122:681-688.
Lakhani P, Kim W, Langlotz CP. Automated extraction of crucial test values and communications from unstructured radiology reports: an analysis of 9.3 million reports from 1990 to 2011. Radiology 2012;265:809-818.
Naik SS, Hanbridge A, Wilson SR. Radiology reports: examining radiologist and clinician preferences regarding style and content. AJR 2001; 176:591-598.
Thomas BJ, Ouellette H, Halpern EF, Rosenthal DI. Automated computer-assisted categorization of radiology reports. AJR 2004; 184:687-690.
Thrall JH. Changing relationships between radiologists and hospitals. Part I. background and major issues. Radiology 2007; 245:633-647.