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

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 radiological 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 interested 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 all keywords to create the RAI and DCC algorithms.
2. 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:
3. Limiting the RAI search to the impression field of the report.
4. Employing sentence level analysis to enhance the accuracy of the algorithm.

RESULTS

For an automated process for detecting RAI and DCC can demonstrate good when using human grading as reference standard. 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 .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 possible to produce in-house RAI and group measurements of radiologist performance metrics. The implications of successful automatic 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 detecting RAI and DCC 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 acknowledgment of the direct indication for the report.

DISCUSSION

REFERENCES

