Performance of a Machine Learning Classifier of Knee MRI Reports in Two Large Academic Practices: A Tool to Estimate Diagnostic Yield

Abstract

*Purpose:* To evaluate the performance of a natural language processor in classifying a database of free text knee MRI reports in two separate academic radiology practices.

*Methods:* A natural language processing (NLP) system that uses terms and patterns in manually classified narrative knee MRI reports was constructed. The NLP was trained and tested on expert-classified knee MRI reports from two major healthcare organizations. Radiology reports were modeled in the training set as vectors and used a support vector machine framework to train the classifier. A separate test set from each organization was used to evaluate the performance of the system. We evaluated the performance of the system both within and across organizations. Standard evaluation metrics such as accuracy, precision, recall and F1 score and their confidence intervals were used to measure the efficacy of our classification system.

*Results:* Accuracy for radiology reports that belong to the model’s clinically significant concept classes after training data from the same institution: F1-score greater than 90% (CI 89.4%-92.9%). Performance of the classifier on cross institutional application without institution-specific training data: F1-scores of 77.6% and 90.2% respectively (CI 76.5%-90.4%).   
*Conclusion:* The results demonstrate excellent accuracy by the machine learning classifier in classifying free text knee MRI reports, supporting the institution-independent reproducibility of knee MRI report classification. Further, the machine learning classifier performed well on free text knee MRI reports from another institution. These data support the feasibility of multi-institutional classification of radiology imaging text reports with a single machine learning classifier without requiring institution specific training data.

Background

Imaging costs remain a significant proportion of health care expenditures. The size and growth of these costs concern policy makers, payers and society alike; diagnostic imaging accounts for 10 percent of total annual healthcare costs, or $100 billion (1-5). While the growth rate of medical imaging utilization has recently decreased (cite), overall MRI and CT utilization rates remain at an all-time high for Medicare beneficiaries (5). Even when imaging is beneficial, the costs associated with current utilization levels are difficult to sustain; thus, studies that analyze imaging appropriateness and evaluate approaches to identify low yield imaging may be essential to the continued health of radiology as a specialty.

One effective strategy used to evaluate imaging utilization and appropriateness has been through the analysis of diagnostic yield as a comparative metric of appropriate imaging service utilization (1-4, 6). This approach relies on analyzing the radiology report to classify the imaging study according to predetermined criteria, a task that typically requires manual expert review, an extremely time consuming and inefficient task even for small datasets. However, the barriers to automating the analysis and classification task are also high, as opposed to analysis of discrete clinical and laboratory data, because classification of the imaging report narrative requires understanding of context, syntax, structure, and terminology. Yet recent advancements in machine learning techniques have shown potential in automated free narrative text analysis via the use of natural language processing (NLP). For example, radiology reports have been successfully analyzed using natural language processing techniques to extract clinically important findings and recommendations (7, 8); NLP may hold the potential to automate the task of classifying imaging reports in a way that could inform decisions regarding medical imaging utilization and appropriateness (1-4, 6).

The purpose of this study is to evaluate the performance of a natural language processor in classifying a database of free text knee MRI reports in two separate large academic radiology practices. By automating the classification of imaging reports, this work could enable automated identification of low-yield imaging tests, thereby identifying targets for the reduction of unnecessary imaging examinations.

**Methods**

This retrospective study was approved by the institutional review board of all participating institutions. A waiver of informed consent was granted.

*Study Design*

We built a natural language processing (NLP) system that uses terms and patterns in manually classified narrative knee MRI reports to identify normal and abnormal cases automatically in new unseen knee MRI reports. We trained and tested the system on expert-classified knee MRI reports from two major healthcare organizations, Group 1 and Group 2. To build our system, we modeled radiology reports in the training set as vectors and used a support vector machine (SVM) framework to train the classifier. A separate test set from each organization was used to evaluate the performance of the system. Standard evaluation metrics such as accuracy, precision, recall and F1 score and their confidence intervals were used to measure the efficacy of our classification system.

We evaluated the performance of the system both within and across organizations. The evaluation tested the generalizability of our system to organizations with no manually labeled training data. We also measured the standard classification metrics and their confidence intervals.

*Knee MRI Report Data Set*

The data set comprised knee MRI reports from two major healthcare organizations, Group 2 and Group 1. The Group 1 data was derived from previously published work to create a manually annotated radiology report database (3, 6, 9). This database contained all knee MRI reports from a 12-month period (2011-2012). The Group 2 data comprised all knee MRI reports from a 24-month period (2008-2010). To control for the confounding effects of prior surgery and post-operative changes, we included only the first-ever knee MRI scans in patients without a prior history of knee surgery. All the eligible reports were annotated by a human expert using previously published criteria (Cite Group 1 paper). In total 706 Knee MRI reports from Group 1 and 1748 Knee MRI reports from Group 2 were manually labeled.

MRI scans with one or more of the following diagnoses were considered positive for the purpose of this study: anterior cruciate ligament (ACL) tear, posterior cruciate ligament (PCL) tear, meniscal tear (MT), medial collateral ligament (MCL) injury, lateral collateral ligament (LCL) injury, patellofemoral extensor mechanism injury (PD), full-thickness cartilage or osseous defect (FTCT) (Table 1). An additional “other” category was included for infrequently encountered findings not otherwise categorized (i.e. avascular necrosis, mass, fractures).

*Natural language processing*

To build our NLP system, we randomly partitioned the radiology reports from each organization into a training set with 80% of the reports and a test set with the remaining 20% of the reports. We used the training set to create an SVM classifier for each organization by using machine learning techniques.

To transform the radiology reports into data that could be processed by an SVM classification framework, we modeled the radiology report text quantitatively. Each radiology report was modeled as a vector, where each vector dimension corresponds to an n-gram. These n-grams are contiguous sequences of one, two or three tokens (words) in the report. If a report contains an n-gram, that n-gram has a non-zero weight in the report’s vector representation. The weight of each n-gram was computed using term frequency–inverse document frequency (tf-idf), a common weighting scheme in text mining (10). A tf-idf weight increases proportionally by the n-gram frequency in the report, and is scaled down by the commonality of the n-gram among all reports in the data set.

An SVM is a maximum-margin classifier, which finds the decision boundary with the largest separation between positive and negative training examples in the training set. In this work we used LIBSVM, a widely-used open source machine learning library, to train our SVM classifier.

The performance of the SVM classifier was initially evaluated on the same organization’s test set. In addition, to explore the NLP system would perform on a new organization’s data where no labeled training set is available, we evaluated each SVM model on the other organization’s data set. In these evaluations we measured accuracy, precision, recall and F1 score for the models on the data sets. We also calculated the confidence intervals for all of the evaluation metrics through bootstrapping (11)

**Results**

Table 1 shows the performance of the SVM classifier and the associated confidence intervals when they are trained and tested on data from the same organization. Table 2 shows the same metrics when the training and testing data come from two different organizations. ROC curves are provided.

***95% confidence intervals for all of the evaluation metrics through the asymptotic approach in R statistical toolbox.***

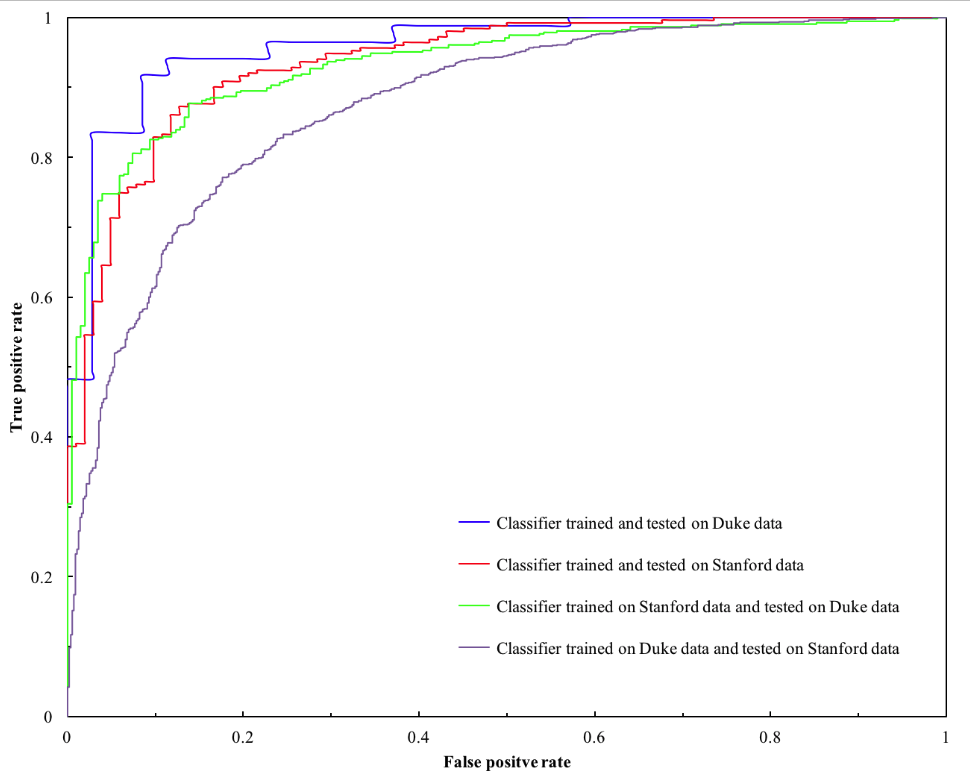
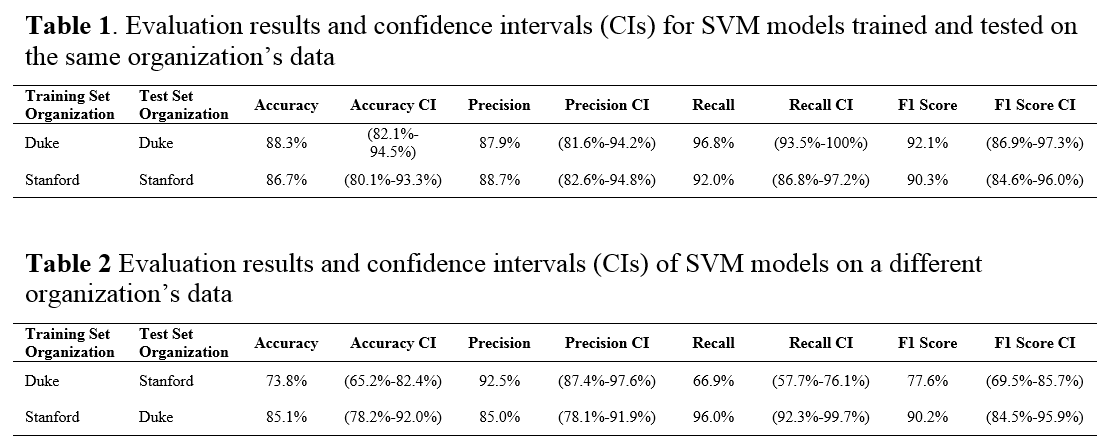


Figure. ROC curves of our four classifiers trained and evaluated on different data sets.

**Discussion**

The purpose of this study was to evaluate performance of a natural language processor (NLP) in automated classification of knee MRI imaging reports from two separate academic institutions. The results demonstrated excellent accuracy for radiology reports that belong to the model’s clinically significant concept classes (F1-scores greater than 90%) supporting the institution-independent reproducibility of knee MRI report classification. Further, the cross institutional results reveal that even without institution-specific training data, the classifier performed well on knee MRI reports from another institution and support broad multi-institutional classification of radiology imaging reports with a single classifier.

A substantial foundation of research over the past two decades has aimed to address automated data extraction from radiology reports; early studies used increasingly sophisticated classification schema (12-16). Well studied early programs include MedLEE (Medical Language Extraction and Encoding System), which extracts and translates narrative radiology reports into predefined controlled vocabulary terms as a preliminary step for additional program analysis (12). More recently, an NLP-based system called SAPHIRE analyzed text with the Unified Medical Language System (UMLS) Metathesaurus to automate indexing of radiology reports for clinical image repositories (16). Other groups have used advanced machine learning techniques to extract more specific information from radiology reports. For example, Sohn et al used machine learning to identify patients with abdominal aortic aneurysms (17), while Dang 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 (18). In contrast to the LEXIMER model, which achieved accuracy of over 95%, our NLP model demonstrates accuracy greater than 85% and F1 statistic greater than 90% for classification tasks. The difference is likely due to the lack of structured reporting in knee MRI reports, as well as diversity and ambiguity of normal and abnormal reports.

One major challenge for NLP classification of medical imaging reports is the broad applicability across institutions, because semantic variations, stylistic differences, and variants of pathologic and anatomic terminology are employed by different radiologists within and between organizations. To the best of our knowledge, this is the first study to analyze the classification performance of an NLP trained on data from one institution and applied to the unstructured data from another. Because the training data effectively “teaches” the NLP the patterns of text and semantic styles at the organization from which the reports are obtain, inter-institutional differences could limit the broad applicability of NLP in classifying reports from another health care institution. We found that inter-institutional classification accuracy, while acceptable, was not as high as intra-institutional analysis. This decrement in accuracy is likely due, in part, to the differences in the radiology documentation methods between the different institutions. For example, the Group 1 radiologists tended to use a uniformly narrow range of language with regard to ACL tear (sprain, partial, full-thickness tear) while the Group 2 radiologists used a wider range of terms to describe ACL injuries (disruption, rupture) which were not found in the Group 1 reports. Nevertheless, despite some differences in regional semantics, terminology, and report format, the program was capable of achieving high (80-90% on average) accuracy.

The increasing availability of computational methods to process vast amounts of unstructured information make it possible to inform clinical decision making directly from large repositories of narrative medical data (19, 20). This approach to information extraction can yield insights which would be virtually impossible to obtain with prospective randomized trials (21). This approach is particularly valuable for radiology clinical decision support to manage imaging utilization. For example, understanding the rate of normal examinations (“negative studies”), can serve as an indirect marker of utilization appropriateness (1-3, 6, 9, 22) Although the narrative structure of radiology reports presents challenges for automated content analysis, our results demonstrate that medical imaging reports can be accurately and automatically classified based on pre-set criteria. These techniques may enhance the ability of imaging report databases to contribute to the development of evidence-based imaging utilization and automated decision support strategies. For example, this classifier could be used as part of a larger data analysis effort by healthcare providers, payers, and/or accountable care organizations to evaluate imaging utilization and target cohorts of inappropriate utilization or overutilization. Further, this classifier could also be used in development of a clinical decision support tool to identify more specific actionable clinical scenarios that result in low-yield imaging tests.

This study has limitations. The machine learning classifier did not achieve perfect accuracy. Yet, the accuracy measures are comparable to other published NLP studies. Our dataset includes radiology reports from large tertiary care centers and the finding trends obtained may not reflect trends from other institutions or in smaller programs. Evaluation of more broad applicability of this system will be required to draw conclusions regarding multi-institutional utility.

In conclusion, we report the first results of a machine learning classifier that can accurately categorize free text knee MRI reports based on pre-defined criteria demonstrating both intra- and cross-institutional fidelity; these data suggest feasibility of more broad usage of natural language processing in automated classification of imaging text reports. Application of these machine learning techniques in order to leverage data found by classifying free text imaging reports may aid healthcare providers, payers, and/or accountable care organizations as part of a larger strategy to optimize imaging utilization appropriateness and reducing overall imaging costs via medical data analysis, even when applied to data originating from multiple institutions.

Acknowledgements: This work was made possible in part by [**BLINDED**] which is funded in part by NIH grant [**BLINDED**].

References

1. Amrhein TJ, Paxton BE, Lungren MP, et al: Physician self-referral and imaging use appropriateness: negative cervical spine MRI frequency as an assessment metric. AJNR. American journal of neuroradiology 2014; 35:2248-2253

2. Lungren MP, Paxton BE,Kilani RK: Imaging self-referral: here we go again. AJR. American journal of roentgenology 2013; 201:W658

3. Lungren MP, Amrhein TJ, Paxton BE, et al: Physician self-referral: frequency of negative findings at MR imaging of the knee as a marker of appropriate utilization. Radiology 2013; 269:810-815

4. Kilani RK, Paxton BE, Stinnett SS, et al: Self-referral in medical imaging: a meta-analysis of the literature. Journal of the American College of Radiology : JACR 2011; 8:469-476

5. Lang K, Huang H, Lee DW, et al: National trends in advanced outpatient diagnostic imaging utilization: an analysis of the medical expenditure panel survey, 2000-2009. BMC medical imaging 2013; 13:40

6. Paxton BE, Lungren MP, Srinivasan RC, et al: Physician self-referral of lumbar spine MRI with comparative analysis of negative study rates as a marker of utilization appropriateness. AJR. American journal of roentgenology 2012; 198:1375-1379

7. Dreyer KJ, Kalra MK, Maher MM, et al: Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology 2005; 234:323-329

8. Yetisgen-Yildiz M, Gunn ML, Xia F, et al: A text processing pipeline to extract recommendations from radiology reports. Journal of biomedical informatics 2013; 46:354-362

9. Amrhein TJ, Lungren MP, Paxton BE, et al: Journal Club: Shoulder MRI utilization: relationship of physician MRI equipment ownership to negative study frequency. AJR. American journal of roentgenology 2013; 201:605-610

10. Manning CD, Raghavan P,Schütze H: Introduction to information retrieval, New York, Cambridge University Press, 2008

11. Efron B,Tibshirani R: An introduction to the bootstrap, New York, Chapman & Hall, 1993

12. Friedman C, Alderson PO, Austin JH, et al: A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association : JAMIA 1994; 1:161-174

13. Jain NL, Knirsch CA, Friedman C, et al: Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. Proceedings : a conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium 1996; 542-546

14. Jain NL,Friedman C: Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proceedings : a conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium 1997; 829-833

15. Hripcsak G, Austin JH, Alderson PO, et al: Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 2002; 224:157-163

16. Hersh W, Mailhot M, Arnott-Smith C, et al: Selective automated indexing of findings and diagnoses in radiology reports. Journal of biomedical informatics 2001; 34:262-273

17. Sohn S, Ye Z, Liu H, et al: Identifying Abdominal Aortic Aneurysm Cases and Controls using Natural Language Processing of Radiology Reports. AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science 2013; 2013:249-253

18. Dreyer KJ, Kalra MK, Maher MM, et al: Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: Validation study. Radiology 2005; 234:323-329

19. Leeper NJ, Bauer-Mehren A, Iyer SV, et al: Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PloS one 2013; 8:e63499

20. Gallego B, Dunn AG,Coiera E: Role of electronic health records in comparative effectiveness research. Journal of comparative effectiveness research 2013; 2:529-532

21. Frankovich J, Longhurst CA,Sutherland SM: Evidence-based medicine in the EMR era. The New England journal of medicine 2011; 365:1758-1759

22. Raja AS, Ip IK, Prevedello LM, et al: Effect of computerized clinical decision support on the use and yield of CT pulmonary angiography in the emergency department. Radiology 2012; 262:468-474