Radiology Diagnostic Exchange Agents: Clinical Term Identification and Validation

Kenneth B. Shores¹, Joseph A. Konstan, PhD.¹, John T. Riedl, PhD.¹, Andrew J. Taylor M.D.³,⁴, Daniel J. Steinberger, M.D.³,⁴, Terrence J. Adam, M.D., PhD.²,³

¹. Computer Science 2. Pharm. Care and Health Systems 3. Institute for Health Informatics 4. School of Medicine.; University of Minnesota, Minneapolis, MN.

Abstract: Diagnostic radiology is one of the key areas of clinical diagnosis on which much of the health care system is built. Along with pathology, radiology has a unique role in providing diagnostic information for prognosis, treatment, and management of clinical conditions. This role is clinically challenging due to the problems of knowledge management associated with the free-text radiology reports which are currently the standard of practice for radiology care. In order to address this critical knowledge management problem, we have proposed a solution using the Radiology Diagnostics Exchange Agent, which is under development and will enhance clinical care management. Using a human computation approach, we have started to identify and validate clinically actionable terms on which an information management infrastructure can be developed with important implications for clinical care and research.

Introduction: The use of clinical metadata that conveys meaning is important for the continued development of better electronic medical records and to support future research. Current use of natural language processing and text mining provide important capabilities for gaining meaning from unstructured clinical text, but rarely interact with or impact the documentation process of providers and often require clinician validation of analytic findings. In order to improve radiology reporting, enhancements in the clinical documentation process by providers are needed to identify the appropriate diagnostically relevant data elements. Standardization of this process is a long-standing radiology problem (Hickey 1922). The identification of key clinical terms which represent actionable events are needed in the clinical record to accurately identify episodes of care, support the development of quality assessment metrics (Sickles 2002) and manage clinical cost metrics. In addition, the identification of high value discrete data elements in the context of clinical reports can be used to enhance intra-provider communications by delivering care guidelines, best practice data, and relevant clinical literature at the point of care.

Conceptual Framework: We define radiology diagnostic exchange agents (RDEAs) as a knowledge management process utilizing clinical terms which convey clinically actionable information on patient diagnosis data in the radiology report with links to clinical problems, temporal data, and contextual information. These terms frequently parallel the list of clinical diagnosis data, but also have additional information elements which impact and inform downstream clinical activities associated with patient treatment and follow-up. These RDEAs are important in managing care decisions because they signify a need for action or careful attention by the care provider. An example would be the finding of a “pulmonary nodule” in a radiology report. Expected actors impacted by the identification of a RDEA concept include the radiologist’s imaging interpretation for process initiation, the referring provider to establish clinical follow-up, the researcher to help enroll patients in clinical studies, the quality improvement officer to manage a follow-up registry and finally the patient who can better track their medical history and clinical care.

Discovering consistent and accurate interpretation of the findings in the radiology report is an information management problem for clinical diagnostics, results management and clinical data aggregation. Obtaining clinical expert review of the actionable terms was our initial step in project work. In order to maximize the efficiency of the reviewer, a graphical interface tool was developed to assist in data curation and designed to support free text review of any clinical text content. The tool was iteratively developed by a team of clinical providers, computer scientists and library scientists with the goal of facilitating a human computation task for experts to identify relevant terms. The design of this tool is notable because it demonstrates an iterative process for developing an appropriate tool for
data aggregation. The data collected by this tool addresses two primary questions. First, how many radiology reports must be reviewed to efficiently find a sufficient quantity of RDEA terms? Second, what are the trade-offs for adding more experts to identify terms? Finally, did provider specialty impact their term assessment?

**Background:** Radiology reporting provides key clinical information in which diagnostic data is acquired from expert interpretation of imaging data. This data may support a critical inflection point in the care pathway with implications for ongoing care. Though radiology has a key role in clinical diagnostics, the information in the radiology report may not directly provide a new diagnosis but can provide valuable supporting data. Radiology reports include a mixture of structured and unstructured data created by the radiologist. In order to assist in the interpretation of results and to maintain proper documentation for referring providers, the radiology report should be standardized to improve the ability to support data management for research and clinical care. (Sobel 1996). In order to improve the quality of radiology reports, the identification of key content is needed for downstream decision support and management of findings in a knowledge management framework such as the RDEA. Natural language processing, text mining and statistical methods provide many tools to identify elements that are relevant in the reporting process. However, in order to identify content for clinical care management it is necessary to have clinical expert review of report content since reports may have a large number of identifiable clinical terms yet only a small number will require active management and long-term follow-up. To address this problem, we designed a human computation system to help clinical experts to quickly identify important terms that should be classified for building the proposed RDEA corpus. Human computation is a common approach used in a variety of domains to avoid the difficulties in developing a fully automated tool to perform a complicated task (Alexander 2011).

**Methods:** The study setting for the development, testing and implementation occurred in a large University-affiliated tertiary clinical care setting. Institutional Review Board approval was obtained prior to study initiation.

**Data Source:** Reports for the task were randomly selected from a pool of 200,000 de-identified radiology reports. Reports were selected based on clinical radiology reporting specialty focusing on body image reports. Of the 400 selected, 356 were computerized tomography reports supplemented with 44 reports from magnetic resonance imaging. The report data set was previewed for protected health information content and the identification and replacement of empty reports. All reports were encoded and stored on an encrypted protected health information server.

**Data Collection:** Key terms were identified in the radiology reports provided to the clinical expert reviewers. Providers represented radiology (4), general internal medicine (2) and pathology (1) including one dual trained clinician. Each reviewer was given instructions and training on the use of the system for term marking. Each subject reviewed the reports in a different randomly selected order to avoid any informational priming. Subjects reviewed the 400 radiology reports in two separate batches of 200 reports. To facilitate the term marking process and aggregate data for RDEA development, we built an in-browser web application to display and collect feedback on reports. Hosted on a secure server, the application authenticated users and tracked many user actions, including logging time spent viewing each report, storing marked terms and collecting optional free text feedback on reports. As the report-writing software from our sample of data was storing reports as HTML-encoded report text, minimal reformatting was required for displaying the reports in a web interface. After term collection, several data-related and content-driven tasks needed to be completed, including the identification of ambiguous clinical terms and the development of a term set of clinical reference data.

The primary functionality of the application was to allow users to identify an important clinical term by highlighting it with their mouse. The term was then displayed as a link, and if a user had selected a term incorrectly, or changed their mind while continuing to view the report, the term could be unselected by clicking the link. Once a term was added, future reports would pre-link the term as the report loaded. This offered the benefit that after processing several reports, a user would already find past important terms pre-highlighted and users were queried to confirm the
The term remained important in the subsequent report. Users were not able to view terms highlighted by other users, so no user’s collected terms were biased by terms selected by a different user.

**Clinical Term Validation:** Each term which was identified by at least four providers was then reviewed by an independent clinical provider to assess if that term had clinical relevance. Additionally, an equal number of terms selected by only one user (sampled from all users) was mixed in with the highly selected terms, to compare the value between terms with a single selection and terms with many selections. Each item was scored into one of four categories: 1) Clinically actionable 2) Novel Clinical Term 3) Ambiguous Term 4) Not/Unlikely Clinically Actionable. Term classifications which were ambiguous were reviewed with a second clinician to assess for term relevance and potential for clinical action. In addition, for terms for which there was a need for additional context in order to assess term relevance a “context dependent” note was added.

**Results:** Each subject was provided the opportunity to review a total of 400 reports by the system. A report counter was provided within each session to give the provider information on the number of reports they had completed. The report marking pattern has been plotted over time in Figure 1 to demonstrate the user marking behavior.

![Figure 1: Provider Reporting Patterns](image)

The provider marking patterns show that the number of terms marked per report decrease as more reports are processed. 1400 terms were collected across six users, with varying degrees of overlap. Only 20 terms were selected by all six users, but a total of 91 terms were agreed upon by at least four of the six users (Table 1).

| Users Agreeing on a Specific Term | 1   | 2   | 3   | 4   | 5   | 6   |
|-----------------------------------|-----|-----|-----|-----|-----|-----|
| Number of Terms                    | 1044| 183 | 82  | 40  | 31  | 20  |

We found that this set of highly-agreed upon terms tended to represent shorter terms or groups of words. The average length of the 91 terms marked four or more times was 16.78 characters, while the terms marked by only one user had an average length of 30.70 characters. An observation of the term sets suggests that terms marked once are more likely to be long, report-specific terms which might contain multiple concepts, and often contained more
context-specific report content. Users who completed all 400 reports spent between 6 and 10 hours, spread across several sessions. We also found a larger number of terms marked by one provider, user_6. This user represented the only pathology trained provider in the reviewer group, providing some evidence that clinical specialization may affect the decision of scoring clinical terms as significant. For comparison, this user marked 1009 terms, where the next highest user marked only 314 terms.

We plot the average rate of term collection in Figure 2 for five of our six users. We exclude user_6 for reasons described above. A power-law curve was fit to the moving average (window size 3) of the user term-marking rate. The fit has an $R^2$ value of .59475. We found a power-law provided the best fit and the result met our expectation from the data. We expect a power law fit of the form $a x^b$ because we expect two factors to logarithmically reduce the likelihood of identifying terms over time. First, as terms are identified, fewer potential terms exist to be selected, which reduces the probability of finding new terms in a new report. Second, we believe that not all terms have the same visibility. Since users were asked to prioritize term selection based on clinical importance, some terms will have lower importance than others. These terms are unlikely to be “important enough” to select until a report is found where no other, more important terms are presented. These terms will tend to become less frequent over time, since new reports are more likely to already contain pre-highlighted terms of higher importance.

![Figure 2: Rate of Term Marking per Report](image)

Following our initial collection of terms, a selection of highly-agreed-on terms and terms that were only highlighted by a single user were chosen for validation by a radiologist. From a set of 198 terms, each term was scored by this expert as being in one of the four categories listed above. The reviewing expert was blinded to the number of times a term was highlighted. For our analysis, we considered terms in condition 1 (clinically actionable) to be terms that fit RDEA criteria. Terms in the other categories (novel, ambiguous or not clinically actionable) were treated as “not clinically actionable,” because both the ambiguous and novel cases are unlikely to require attention and have limited literature or best practices data available. An example ambiguous term, selected only once, was “size criteria.” This provides no clear or meaningful information and is unlikely to be useful without context. An analysis was performed on the term sets to assess the likelihood of being clinically actionable, dependent on the number of times a term was selected. A chi squared test was performed on the similarity of the selection groups which determined that the groups were statistically significantly different ($p<.009$) with a 95% confidence interval of (-0.32, -0.05). Terms with clinical significance were more likely to be selected by multiple providers.
**Discussion:** The term marking results supported the expected pattern that the number of marked terms would decrease as providers reviewed more reports. Though not all providers were able to complete the 400 report set, there was a demonstrated rate of reduced term marking. To completely map the vocabulary of the term space we set out to investigate, a large amount of reports would be processed such that no new terms would be added. However, the goal of the project was not to identify all fitting terms and instead to explore a mechanism for discovering these terms efficiently, and to get a large set of appropriate terms to seed future work. Several findings in our work provide useful insights for future work on radiology reports for both research and clinical practice. The results provide data on a feasible case study for the application of human computation to a complex task in a medical setting. In addition, our results indicate our approach is an effective and efficient method for data aggregation which can be implemented quickly and easily, supporting multiple simultaneous users. Though using a physician review to identify terms uses high-cost provider review time, it is has clear clinical relevance and is more efficient and reliable than attempts to add extra work to normal clinician workflow, such as logging terms as discovered when viewing a report in clinic. In addition, this work provides important preliminary data for RDEA development and shows feasibility of the data collection and validation process.

The study has several limitations which may impact the generalizability of the results. The clinical reviewers were limited to physicians with radiology, internal medicine, and pathology backgrounds. Having a larger group of reviewers with additional clinical specialty representation may be needed to optimally identify clinical terms. Furthermore, since the pathology physician demonstrated a different marking pattern there may be a need to expand on these preliminary findings. The evaluation was completed at a University-affiliated tertiary care clinical site which may not be generalizable to other clinical settings. The context for the study was on body imaging which is one area of radiology practice, providing a narrower project scope to ensure feasibility. It can be assumed that other areas of radiology practice would yield a similar result, however, this has not yet been established.

Future work should include an expansion of expert human-computation tasks as applied to a variety of informatics systems where NLP and text mining are currently insufficient to obtain quality results. Our research group intends to build on the results here in other specialties besides body imaging. We intend to use this term set to seed a collaborative filtering system which associates terms with information resources and builds on the proposed concept of a RDEA. In this work we presented RDEA, a knowledge management approach to support clinical decision-making and radiology report aggregation for research and quality initiatives. We discussed in detail one component of RDEA, clinically interesting terms which are actionable. We developed a system which utilizes human computation to quickly and efficiently leverage clinician and radiologist expertise to identify terms. We discussed validation measures used to assess term quality and accuracy and show that multi-user agreement is more likely to identify clinically actionable terms than any single user. Finally we discuss future work to apply the collected terms for radiology report enhancement.

**References:**

1) Alexander J. Quinn and Benjamin B. Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (CHI ’11). ACM, New York, NY, USA, 1403-1412. DOI=10.1145/1978942.1979148 [http://doi.acm.org/10.1145/1978942.1979148](http://doi.acm.org/10.1145/1978942.1979148)

2) Hickey PM. Standardization of roentgen-ray reports. AM J Roentgenol. 1922

3) Sickles, EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002; 224:861-869.

4) Sobel J, Pearson M, Gross K, et al. Information content and clarity of radiologists' reports for chest radiography Acad Radiol. 1996;3:709-717.