

A Comparison of Two Approaches to Text Processing: Facilitating Chart Reviews of Radiology Reports in Electronic Medical Records

by Julie A. Womack, CNM, APRN, PhD; Matthew Scotch, MPH, PhD; Cynthia Gibert, MD; Wendy Chapman, PhD; Michael Yin, MD; Amy C. Justice, MD, PhD; Cynthia Brandt, MD, MPH

Abstract

Chart review is central to health services research. Text processing, which analyzes free-text fields through automated methods, can facilitate this process. We compared precision and accuracy of NegEx and SQLServer 2008 Free-Text Search in identifying acute fractures in radiology reports.

The term “fracture” was included in 23,595 radiology reports from the Veterans Aging Cohort Study. Four hundred reports were randomly selected and manually reviewed for acute fractures to establish a gold standard. Reports were then processed by SQLServer and NegEx. Results were compared to the gold standard to determine accuracy, precision, recall, and *F*-statistic.

NegEx and the gold standard identified acute fractures in 13 reports. SQLServer identified 2 in a report-based analysis (precision: 1.00; accuracy: 0.97; recall: 0.15; *F*-statistic: 0.26), and 12 in a sentence-by-sentence analysis (precision: 1.00; recall: 0.92; accuracy: 0.92; *F*-statistic: 0.96).

Text-processing tools utilizing basic database or programming skills are comparable, precise, and accurate in identifying reports for review.

Introduction

Chart reviews are of central importance for clinical, epidemiological, and health services research and are the core of some quality improvement initiatives. Manual chart reviews require significant investments of time and personnel. Text processing, which analyzes free-text fields through automated methods, can facilitate this process.

Chart reviews require two distinct processes: identifying appropriate charts, and understanding and extracting the necessary information. The latter process is complex, requiring significant health-related knowledge and the ability to infer and deduce information that is not explicitly provided in the text. Complex tools exist to locate specific terms within a context that describes, for example, whether or not the condition exists or was negated, the anatomic location, whether this is a current or historical finding,

whether the condition belongs to the patient or to someone else (e.g., family history), and the severity of the condition.¹⁻⁵

Less complex tools may be adequate for the first task and would thus be accessible to noninformaticians. The current study compares the accuracy and efficiency of two text-processing tools: SQLServer 2008 Full-Text Search (SQLServer) and NegEx. Our use case is the presence or absence of acute fractures in free-text radiology reports. We then discuss the results in the context of the larger spectrum of health services research.

Background

The Veterans Aging Cohort Study (VACS) is a prospective, observational cohort study of HIV-infected and age-, race-, and site-matched uninfected Veterans who receive their care from Veterans Health Administration (VHA) facilities in the United States. The VACS has consented and enrolled 7,091 patients (3,569 HIV-infected and 3,522 uninfected) in care in Atlanta, Baltimore, the Bronx, Houston, Los Angeles, Manhattan/Brooklyn, Pittsburgh, and Washington, DC.⁶ VACS has access to complete electronic medical record (EMR) data for these individuals.

Methods

This project was implemented to support a case-control study investigating the impact of HIV status on wrist, hip, and vertebral fractures in VACS. The study involved the review of more than 151,270 electronic radiology reports for identification of fractures. The term “fracture” was included in 23,595 reports (16 percent). The authors felt that manual review of 400 reports would be feasible. A power analysis was performed to identify the minimum number of charts that would provide adequate power. Assuming that the proportion of observed successes (that is, both sources agree on the existence of a trait) is 0.10, a large-sample 0.05-level two-sided test of the null hypothesis that kappa is 0.70 will have 80 percent power to detect an alternative kappa of 0.90 when the sample size is 273.⁷

The authors sought to facilitate this effort using text-processing solutions. Two commonly used but different approaches to text processing were evaluated. SQLServer uses structured query language (SQL), which would be familiar to clinicians with database experience, whereas NegEx would be more familiar to those whose background includes computer programming.

Negation is an important contextual feature of information described in clinical reports. A negation is a statement that is a refusal or denial of another statement. Chapman and colleagues estimated that almost half of the conditions indexed in dictated reports are negated and that nearly two-thirds of findings in radiology reports are negated.⁸ Including an assessment of negation status contributes significantly to classification accuracy.⁹ In the context of the current analysis, negations of interest included but were not limited to “no fracture,” “no evidence of fracture,” and “absence of fracture.” Both NegEx and SQLServer can identify and exclude reports in which the diagnosis of interest is negated.

Radiology Reports

Radiology reports are among the most standardized of clinical reports and thus good candidates for an automated review.¹⁰ The radiology reports used in this study followed a general format: the patient’s name, Social Security number, exam date, procedure, the clinical history (completed by referring provider), the report (which included information on anatomical focus and the procedure to be performed), the findings, and the impression or assessment of the radiologist (Table 1).

Determination of Gold Standard

The gold standard was established by two primary care clinicians (J.W. and C.B.). Each separately reviewed the 400 radiology reports and determined whether or not the reports included information on an acute fracture. Acute fractures were defined as those for which there was evidence that the fracture was new, for example, “rule out acute process” in the clinical history section and “There is what appears to be

Smith's fracture affecting the left wrist" in the report section (Table 1). The separate evaluations were then compared to determine agreement. Arbitration for any disagreements was available through a diagnostic radiologist from a local hospital.

Text-Processing Algorithms: SQLServer 2008

We selected SQLServer 2008 Full-Text Search because it is an off-the-shelf solution for simple text-processing tasks that is part of the Microsoft SQLServer 2008 package. It can address negation. SQLServer allows users to issue full-text queries against character-based data in SQLServer tables and then presents rows (radiology reports) that contain the specified information. Free-text notes must first be imported into a table or tables in a SQLServer database. How notes are stored is determined by the data team to suit the needs of the users. For this study, text notes were imported as a complete report, as described earlier, matched with the individual's Social Security number, a unique identifier for the text, the individual's study identifier, and the exam date.

The full-text search engine builds an inverted, stacked, compressed index structure based on individual tokens in the text. This index allows rapid computation of full-text queries that search for rows with particular words or combinations of words. Indexes are located within full-text catalogs, a logical concept that refers to a group of full-text indexes. The "CONTAINS" predicate can be used to search columns for precise or fuzzy matches to words or phrases, as well as to perform proximity (using the term "near") or weighted matches. The "FREETEXT" predicate breaks search phrases up into separate words, stems words to facilitate matching words with different inflections, and then identifies a list of expansions or replacements for terms based on matches to a thesaurus. SQLServer also includes a modifiable stopword list that includes words, such as "a" or "and," that can be excluded from most searches. Excluding these words enhances the speed with which the queries can be processed. Of note, negations are often included in the stopword list.

Identifying Radiology Reports Using SQLServer

The full-text stopword list was modified by excluding common negation terms so that these terms could be included in the search statement (Figure 1). The full-text predicate "CONTAINS" was used to identify acute fractures by finding notes that contained the word "fracture" or some version thereof ("fractu*"), and excluding notes where the term "fractu*" was associated with (that is, was located near) the terms "no," "without," "*not," "r/o," "fail*," "old," "previous," "prior," "healed," and "possible." The terms "no," "without," "*not," "r/o," and "fail*" excluded reports where fractures were negated. The terms "old," "previous," "prior," and "healed" excluded reports where fractures were old, and "possible" excluded reports where fractures were uncertain.

Using SQLServer, we analyzed reports for concepts consistent with "acute fracture." As these results were quite poor, we conducted a post hoc analysis where the reports were reimported as a series of individual sentences with the patient study number, report date, and sentence sequence number to uniquely identify the sentences. These sentences were then analyzed independently using the same code that was used to analyze the reports.

Text-Processing Algorithms: NegEx

NegEx, developed by Bridewell, Chapman, and colleagues, is a text-processing algorithm that was originally developed for use with discharge summaries.¹¹ NegEx input is indexed words or phrases of interest to the user, often in text or Word format, and the output is whether the indexed phrases are negated, affirmed, or being considered. An example of a NegEx markup follows, where underlining indicates a negation phrase, bold indicates the concepts that are negated, and italics indicate concepts that are not negated:

Report: There is no evidence of recent **fracture, dislocation**, or other bone or *joint abnormality*. The soft tissue structures are normal.

Impression: Essentially normal foot.

A Python release of NegEx also performs the indexing using string matches to a default Unified Medical Language System (UMLS) dictionary or to another dictionary that the user either provides or builds with a function included in the release.

The markup example above correctly indicates that fracture and dislocation have been negated by the phrase “no evidence.” However, NegEx does not correctly identify that “joint abnormality” should also be negated, because in the Python version of NegEx that we applied, a negation phrase extends its scope for six terms. The term “joint abnormality” is too remote from the negation phrase to be identified as negated.

Identifying Radiology Reports Using NegEx

The reports were analyzed using NegEx, which was modified to evaluate only the “Diagnosis” and “Report” sections, thus decreasing the risk of false positives. To identify fractures in the report, we used the UMLS concept unique identifier (CUI) “C0016658,” the general CUI for “fracture.” (In the UMLS, meanings [or concepts] associated with words and terms are enumerated by CUIs. For example, the word “culture” has two meanings, as in “Western culture” and “bacterial culture.” While the word is the same, each meaning has a different CUI.) For completeness, we searched the National Library of Medicine’s UMLS Knowledge Source Server to identify all descendant CUIs of this concept. We organized the CUIs in a flat file, removed all duplicates, and identified 113 fracture-related concepts.¹² We included these CUIs with the parent CUI in the dictionary to identify reports that presented information on fractures. We then identified fracture expressions that were negated, using the list of negation triggers included with the NegEx software. In order to capture fractures that were old, we updated the negation list to include the words “prior,” “old,” and “previous” in relation to the word “fracture,” thus using these words as forms of negation for “acute” fractures.

Analysis

Agreement between the two reviewers was assessed by Cohen’s kappa coefficient (κ), a statistical measure of agreement between two raters for categorical items. It is more robust than percent agreement calculation since Cohen’s kappa takes into account the agreement occurring by chance.¹³

Text-processing results were compared to the gold standard to determine accuracy, precision, recall, and the *F*-statistic (Table 2). These measures rely on comparing the results obtained from the test cases (using SQLServer and NegEx) to those obtained by manual chart review (the gold standard). In the current analysis, “true positives” were radiology reports that were found to contain information on acute fractures both by SQLServer or NegEx and by manual chart review. “True negatives” were radiology reports that were found not to contain information on acute fractures by SQLServer or NegEx and also by manual review. “False positives” were reports found to contain information on acute fractures by SQLServer and/or NegEx but not by manual review, and “false negatives” were reports that were found not to contain information on acute fractures by SQLServer and/or NegEx but were found to contain this information by manual review.

Accuracy is a statistical measure of how well a binary classification test correctly identifies or excludes a condition. Precision and recall are measures for evaluating the performance of information retrieval systems.¹⁴ Precision is the proportion of documents retrieved that are relevant to the user’s needs compared to all documents, relevant or irrelevant, that were retrieved by the method being evaluated. It is analogous to positive predictive value. Recall is the proportion of documents that are relevant to the query and that are successfully retrieved compared to all reports determined to be relevant by the gold standard. It is analogous to sensitivity. The *F*-statistic considers both the precision and the recall of the test and can be interpreted as a weighted average of the precision and recall. In this analysis, both precision and recall were weighted the same (Table 2).

Error Analysis

Successful application of text-processing tools to a specific group of clinical notes is an iterative process in which error analysis plays a central role. For both SQLServer and NegEx, initial programming statements were written and applied to the radiology reports. An error analysis was performed on the results to determine where incorrect classifications occurred and the reasons for the incorrect

classifications. Code was then modified to address these errors and then applied to the radiology reports. These iterated modifications were a central aspect of the analysis and continued until no more modifications could be made to the code to improve the outcomes.

Results

The gold standard identified 13 of the 400 radiology reports as including terms for acute fracture. There was 100 percent agreement ($\kappa = 1.0$) between the two clinician reviewers.

SQLServer, when analyzing each report in its entirety, correctly identified two reports as addressing acute fractures. No reports were identified as positive by SQLServer and negative by the manual review (false positives). Eleven reports were incorrectly classified as not including acute fracture information (false negatives), and 387 reports were correctly classified as not containing acute fracture information. Precision (1.00) and accuracy (0.97) were high, but recall and the *F*-statistic were extremely low (0.15 and 0.26, respectively) (Table 2). NegEx correctly identified 13 positive reports and 387 negative reports. There were no false positives or false negatives. The outcome measures were perfect (precision: 1.00; recall: 1.00; accuracy: 1.00; *F*-statistic: 1.00).

In the post hoc sentence-by-sentence analysis, SQLServer correctly identified 12 radiology reports that included terms consistent with acute fracture and 387 reports that did not contain acute fracture information. There were no false positives and one false negative. The outcome measures for the sentence-by-sentence approach improved dramatically (precision: 1.00; recall: 0.92; accuracy: 0.92; *F*-statistic: 0.96).

Error Analysis

In the error analysis, the SQLServer false negative was the only error that could not be eliminated (Table 3). This error occurred because the proximity term “near” does not distinguish between the relative positions of words. A negation term typically precedes the word or concept that it negates. However, in SQLServer, a word is considered to be “near” another regardless of the relative positions of the negation and the word of interest. Therefore, in the sentence “Proximal left fibular fracture also not appreciably changed,” the negation term “not” was “near” fracture, but it negates a subsequent word, “changed,” and not the word “fracture.”

All other errors were addressed. For example (Table 3), the SQLServer false positive occurred because the abbreviation “r/o” for “rule out” was not initially included in our negation list. Once it was added, this false positive was resolved. The NegEx errors were resolved by adding the words “neither,” “fails to,” and “possible” to the negation dictionary.

Discussion

In our sample of 400 radiology reports, we found that SQLServer, when analyzing sentences, and NegEx, when analyzing reports, were precise and accurate in identifying acute fractures. The success of the sentence-by-sentence analysis would be strengthened if it could be repeated as an initial hypothesis and not as a post hoc test. However, even though NegEx achieved 100 percent specificity in this analysis, the argument should not be made that it is the gold standard for identifying charts for chart review. The usefulness of any text-processing tool depends on the user’s skill and familiarity with the tools.

Although our results may be generalizable to searches for acute fractures in radiology reports from different institutions, they cannot be generalized to different report types or to searches for concepts other than acute fractures. As demonstrated by Harkema and colleagues, report types differ substantially across a wide variety of characteristics, all of which may influence the utility and applicability of text-processing tools.¹⁵

In addition, our study is limited because we restricted our analysis to reports that included the word “fracture.” Reports that describe fracture in different or more subtle ways were not included; thus, we cannot judge how well either SQLServer or NegEx would have worked for these alternative definitions.

This limitation would likely not affect a large number of charts given the uniform nature of charting in radiology reports; however, it doubtless omitted a few.

NegEx and the newer version, ConText (see <http://code.google.com/p/negex/>), have been used extensively in studies of negation tools in a wide variety of clinical notes.¹⁶ The current version of SQLServer is newer and has not been used as extensively in clinical research. Further research is necessary to explore the use of SQLServer, as well as NegEx and ConText, on different types of clinical notes. Some of the most challenging are those that have the least standardization, such as clinical progress notes.

Further research must also address users of these technologies, not just the technologies themselves. Text-processing, natural-language processing, and text-mining tools are developed by those with expertise in computer science and informatics. Environments can facilitate the use of many of these tools; however, evaluation of how effectively these tools can be used by individuals without informatics backgrounds is limited, much as with information retrieval research.^{17, 18}

Chart reviews are of central importance to health services research and to healthcare management. Clinicians often need to review charts to evaluate patient outcomes. Validation of diagnostic codes by electronic medical record review and the development of algorithms to enhance accuracy are essential for medical record specialists and researchers alike. Quality improvement initiatives entail identifying procedures or diagnoses that are not always coded. Reviewing 151,000 or even 23,000 radiology reports is not feasible. Automating the chart selection process results in a far smaller sample and reduces the time and personnel required for the review. But tools that can facilitate this process must be accessible to noninformaticians, as informaticians are not typically the ones doing this research. SQLServer and NegEx demonstrate that good text-processing tools can provide excellent outcomes for appropriate tasks when matched to users with appropriate skills.

Julie A. Womack, CNM, APRN, PhD, is an informatics postdoctoral fellow at the Veterans Health Administration Connecticut in West Haven, CT.

Matthew Scotch, MPH, PhD, is an assistant professor of informatics at Arizona State University in Tempe, AZ.

Cynthia Gibert, MD, MSc, is a professor of medicine at the George Washington University School of Medicine and Health Sciences and an attending physician at the Washington DC Veterans Affairs Medical Center.

Wendy Chapman, PhD, is an assistant professor of informatics at University of Pittsburgh, Pittsburgh, PA.

Michael Yin, MD, is an assistant professor in clinical medicine at the Columbia University Medical Center, NY.

Amy C. Justice, MD, PhD, is a professor at the Yale School of Medicine and a Section Chief at Veterans Health Administration Connecticut in West Haven, CT.

Cynthia Brandt, MD, MPH, is an associate professor at Yale Center of Medical Informatics and a staff physician at Veterans Health Administration Connecticut in West Haven, CT.

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs.

Funding for this article was provided by the Information Research and Development Medical Informatics Fellowship Program of the VA Office of Academic Affiliations. The Veterans Aging Cohort Study is funded by the National Institute on Alcohol Abuse and Alcoholism (U10 AA 13566) and the VHA Public Health Strategic Health Core Group.

Notes

1. Friedman, C., et al. "A General Natural-Language Text Processor for Clinical Radiology." *Journal of Women's Health* 1, no. 2 (1994): 161–74.
2. Christensen, L., P. J. Haug, and M. Fiszman. "MPLUS: A Probabilistic Medical Language Understanding System." *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Association for Computational Linguistics*, Philadelphia, July 2002.
3. Chapman, W. W., et al. "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries." *Journal of Biomedical Informatics* 34, no. 5 (2001): 301–10.
4. Harkema, H., et al. "ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports." *Journal of Biomedical Informatics* 42, no. 5 (2009): 839–51.
5. Mutalik, P. G., A. Deshpande, and P. M. Nadkarni. "Use of General-Purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study Using the UMLS." *Journal of the American Medical Informatics Association* 8, no. 6 (2001): 598–609.
6. Justice, A. C., et al. "Veterans Aging Cohort Study (VACS): Overview and Description." *Medical Care* 44, no. 8, suppl. 2 (2006): S13–S24.
7. Sim, J., and C. C. Wright. "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements." *Physical Therapy* 85, no. 3 (2005): 257–68.
8. Chapman, W. W., W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. "Evaluation of Negation Phrases in Narrative Clinical Reports." *Proceedings of the AMIA Symposium* (2001): 105–9.
9. Chu, D., J. N. Dowling, and W. W. Chapman. 2006. "Evaluating the Effectiveness of Four Contextual Features in Classifying Annotated Clinical Conditions in Emergency Department Reports." *AMIA Annual Symposium Proceedings* (2006): 141–45.
10. Chapman, W. W., W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. "Evaluation of Negation Phrases in Narrative Clinical Reports."
11. Ibid.
12. National Library of Medicine. 2008. *UMLS Knowledge Source Server*. Available from <https://login.nlm.nih.gov/cas/login?service=http://umlsks.nlm.nih.gov/uPortal/Login> (accessed July 21, 2009).
13. Cohen, J. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 120, no. 1 (1960): 37–46.
14. Taylor, J. R. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. Sausalito, CA: University Science Books, 1999.
15. Harkema, H., et al. "ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports."
16. Ibid.
17. Hersh, W. R., and D. H. Hickam. "How Well Do Physicians Use Electronic Information Retrieval Systems? A Framework for Investigation and Systematic Review." *Journal of the American Medical Association* 280, no. 15 (1998): 1347–52.
18. Hersh, W. R., et al. "Factors Associated with Successful Answering of Clinical Questions Using an Information Retrieval System." *Bulletin of the Medical Library Association* 88, no. 4 (2000): 323–31.

Table 1

Examples of Radiology Reports

Type of Report	Radiology Report
Acute fracture	<p>Clinical History: r/o acute process</p> <p>Report: LEFT WRIST AND LEFT FOREARM:</p> <p>FINDINGS:</p> <p>Three standard views of left wrist and two views of left forearm were obtained. The latter consists of lateral and oblique views without straight AP view.</p> <p>Impression:</p> <ol style="list-style-type: none"> 1. There is what appears to be Smith’s fracture affecting the left wrist. It consists of impacted comminuted fracture of distal left radial metaphysis having dorsal angulation at the fracture site as well as avulsion fracture of ulnar styloid process. The fracture affects neighboring radiocarpal as well as inferior radioulnar articulations. 2. Small cortical avulsion fracture is also noted on lateral view of wrist at carpo-metacarpal articulation, probably at 3rd carpometacarpal articulation. <p>NOTE: Presence of an accessory ossicle dorsally opposing the triquetrum or lunate.</p> <ol style="list-style-type: none"> 3. Low-grade osteoarthritis at left 1st carpo-metacarpal articulation as well as intermetacarpal articulation between 1st and 2nd. 4. The rest of radius and ulna are normal.
Fracture negation	<p>Procedure: KNEE 3 VIEWS</p> <p>Clinical History: S/p fall from ladder with right hip fracture. Since then also with right knee swelling and knee that “gives out”. Prior films showed effusion and narrowed joint space. Appreciate re-evaluation.</p> <p>Report: RIGHT KNEE THREE VIEWS</p> <p>Impression:</p> <p>There is no acute fracture or dislocation. Mild patchy osteopenia. Probable small volume suprapatellar joint effusion. Mild patellar spurring.</p>
Negation of acute fracture by age of fracture	<p>Procedure: CALCANEUS 2 VIEWS</p> <p>Clinical History: SP FX CALCANEUS</p> <p>Report: See impression.</p> <p>Impression:</p> <p>Compared to the prior study of 1/12/99 there is no significant change in the deformity of the calcaneus consistent with an old fracture which has undergone complete healing. The adjacent bony structures appear intact.</p>
Implicit negation	<p>Procedure: FOOT 3 OR MORE VIEWS</p> <p>Clinical History: 49 yo b m w cc of bilateral foot pain. please eval thanks.</p> <p>Report: See Impression.</p> <p>Impression:</p> <p>BILATERAL FEET: No traumatic, arthritic, or developmental abnormality is discernible.</p>

Table 2

Precision, Recall, Accuracy, and F -statistic for SQLServer (Sentence), SQLServer (Report), and NegEx

		Fracture Status as Determined by Gold Standard		
		Positive	Negative	
Fractures identified by SQLServer (report)	Positive	2 reports ^a	0 reports ^b	Precision: 1.00
	Negative	11 reports ^c	387 reports ^d	Negative predictive value: 1.00
		Recall: 0.15	Specificity: 1.00	Accuracy: 0.97 F-statistic: 0.26
Fractures identified by SQLServer (sentence)	Positive	12 reports ^a	0 report ^b	Precision: 1.00
	Negative	1 reports ^c	387 reports ^d	Negative predictive value: 0.99
		Recall: 0.92	Specificity: 1.00	Accuracy: 0.92 F-statistic: 0.96
Fractures identified by NegEx	Positive	13 reports ^a	0 reports ^b	Precision: 1.00
	Negative	0 reports ^c	387 reports ^d	Negative predictive value: 1.00
		Recall: 1.00	Specificity: 1.00	Accuracy: 1.00 F-statistic: 1.00

^a true positives

^b false positives

^c false negatives

^d true negatives

Notes:

Accuracy: $(\text{true positives} + \text{true negatives}) / (\text{true positives} + \text{true negatives} + \text{false positives} + \text{false negatives})$

Precision: $(\text{true positives}) / (\text{true positives} + \text{false positives})$

Recall: $(\text{true positives}) / (\text{true positives} + \text{false negatives})$

Negative predictive value: $(\text{true negatives}) / (\text{true negatives} + \text{false negatives})$

Specificity: $(\text{true negatives}) / (\text{true negatives} + \text{false positives})$

F -statistic: $2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$

Table 3

Samples from Error Analyses

Tool	False Positive	False Negative
SQLServer	“r/o fracture”	“Orthopedic external stabilization device through the proximal and mid tibia and fibula with comminuted fractures. . . . Proximal left fibular fracture also not appreciably changed. Left ankle soft tissue swelling. Man-to-man left ankle degenerative change. Overlying opaque material distal left fibular deformity on lateral view without acute fracture line visualized. Mild bony demineralization suspected.”
NegEx	“Neither side shows fracture” “The study fails to demonstrate evidence of acute fracture or dislocation.”	“Possible stress fracture of 1st metatarsal bone.”

Figure 1

Process of Radiology Report Analysis with SQLServer and NegEx

