Original Investigation

Comparison of Natural Language Processing Rules-based and Machine-learning Systems to Identify Lumbar Spine Imaging Findings Related to Low Back Pain

W. Katherine Tan, BS, Saeed Hassanpour, PhD, Patrick J. Heagerty, PhD, Sean D. Rundell, DPT, PhD, Pradeep Suri, MD, MS, Hannu T. Huhdanpaa, MD, MSc, Kathryn James, PA, MPH, David S. Carrell, PhD, Curtis P. Langlotz, MD PhD, Nancy L. Organ, BA, Eric N. Meier, MS, Karen J. Sherman, PhD, MPH, David F. Kallmes, MD, Patrick H. Luetmer, MD, Brent Griffith, MD, David R. Nerenz, PhD, Jeffrey G. Jarvik, MD, MPH

Rationale and Objectives: To evaluate a natural language processing (NLP) system built with open-source tools for identification of lumbar spine imaging findings related to low back pain on magnetic resonance and x-ray radiology reports from four health systems.

Materials and Methods: We used a limited data set (de-identified except for dates) sampled from lumbar spine imaging reports of a prospectively assembled cohort of adults. From N = 178,333 reports, we randomly selected N = 871 to form a reference-standard dataset, consisting of N = 413 x-ray reports and N = 458 MR reports. Using standardized criteria, four spine experts annotated the presence of 26 findings, where 71 reports were annotated by all four experts and 800 were each annotated by two experts. We calculated inter-rater agreement and finding prevalence from annotated data. We randomly split the annotated data into development (80%) and testing (20%) sets. We developed an NLP system from both rule-based and machine-learned models. We validated the system using accuracy metrics such as sensitivity, specificity, and area under the receiver operating characteristic curve (AUC).

Results: The multirater annotated dataset achieved inter-rater agreement of Cohen's kappa > 0.60 (substantial agreement) for 25 of 26 findings, with finding prevalence ranging from 3% to 89%. In the testing sample, rule-based and machine-learned predictions both had comparable average specificity (0.97 and 0.95, respectively). The machine-learned approach had a higher average sensitivity (0.94, compared to 0.83 for rules-based), and a higher overall AUC (0.98, compared to 0.90 for rules-based).

Conclusions: Our NLP system performed well in identifying the 26 lumbar spine findings, as benchmarked by reference-standard annotation by medical experts. Machine-learned models provided substantial gains in model sensitivity with slight loss of specificity, and overall higher AUC.

Key Words: Natural language processing; lumbar spine diagnostic imaging; low back pain.

Acad Radiol 2018; ■:■■–■■

From the Department of Biostatistics (W.K.T., P.J.H., N.L.O., E.N.M.); Center for Biomedical Statistics, University of Washington, Seattle Washington (W.K.T., P.J.H., N.L.O., E.N.M.); Department of Biomedical Data Science, Dartmouth College, Hanover, New Hampshire (S.H.); Department of Health Services, University of Washington, Box 357660, Seattle WA 98195-7660 (S.D.R., J.G.J.); Department of Rehabilitation Medicine, University of Washington, Seattle (S.D.R., P.S.); Division of Rehabilitation Care Services, Seattle Epidemiologic Research and Information Center, VA Puget Sound Health Care System, Seattle, (P.S.); Radia, Inc. 19020, Lynwood, Washington (H.T.H.); Department of Radiology, University of Washington, 1959 NE Pacific Street, Seattle WA 98195 (K.J., J.G.J.); Kaiser Permanente Washington Health Research Institute, Seattle, Washington (D.S.C., K.J.S.); Department of Radiology, Stanford University, Palo Alto, California (C.P.L.); Department of Radiology Mayo Clinic, Rochester, Minnesota (D.F.K., P.H.L.); Department of Radiology (B.G.); Neuroscience Institute, Henry Ford Hospital, Detroit, Michigan (D.R.N.); Department of Neurological Surgery, University of Washington, 1959 NE Pacific Street, Seattle, WA 98195 (J.G.J.); Comparative Effectiveness, Cost and Outcomes Research Center, University of Washington, 4333 Brooklyn Ave NE, Seattle, WA 98105 (S.D.R., P.S., K.J., J.G.J.). Received January 9, 2018; revised March 9, 2018; accepted March 9, 2018. **Address correspondence to:** J.G.J. e-mail: jarvikj@uw.edu

https://doi.org/10.1016/j.acra.2018.03.008

INTRODUCTION

Low back pain (LBP) has an estimated global lifetime prevalence of almost 40% (1). In the United States, LBP is the second most common symptom prompting physician visits (after respiratory infections), with an estimated annual cost of over $100 billion (2,3). Despite numerous available interventions for this common and burdensome condition, LBP remains difficult to diagnose and to treat effectively (4). One difficulty in addressing LBP is substantial heterogeneity in its etiology, progression, and response to treatment. For instance, a clinical presentation of LBP could be caused by reasons ranging from minor muscle strains to malignant tumor (5,6).

The discovery of patient subgroups with similar prognoses and intervention recommendations is a research priority for advancing LBP care (7,8). Spine imaging findings may help define such subgroups. In most cases, imaging findings alone are insufficient to diagnose the underlying reasons for LBP. Furthermore, even when present, imaging findings are often of uncertain clinical significance given their frequent presence in asymptomatic individuals (9). Yet, certain imaging findings, such as endplate changes, are more prevalent in patients with LBP compared to nonclinical populations (10). To understand relationships between imaging findings and LBP, an important step is the accurate extraction of findings, such as stenosis and disc herniation, from large patient cohorts.

Radiologists identify lumbar spine imaging findings on images and create reports containing these findings. Although information extraction from these reports can be done manually, this technique is impractical for large sample sizes. As an alternative to manual extraction, natural language processing (NLP) has been successfully used to harvest specific findings and conditions from unstructured radiology reports with high accuracy. For example, a model to identify pulmonary nodules from computed tomography reports attained a positive predictive value (PPV) of 0.87 (11). Another group achieved an average specificity of 0.99 applying complex automated queries to identify 24 conditions from chest x-ray reports, including neoplasms, pneumonia, and tuberculosis (12). Such methods have not been previously applied to lumbar spine degenerative findings commonly found in patients with LBP. Automated identification of these findings is an important step in building clinical information systems that can support large-scale learning approaches to improve both clinical care and clinical research.

In this manuscript, we describe the development and evaluation of an NLP system for identification of 26 lumbar spine imaging findings related to LBP on magnetic resonance (MR) and x-ray radiology reports. Our set of 26 imaging findings includes eight findings commonly found in subjects without LBP, as well as additional findings that are less common but are potentially clinically important (9).

MATERIALS AND METHODS

Reference-standard Dataset

We used a limited data set (de-identified except for dates of service) and our study protocol was deemed minimal risk with waivers for both consent and Health Insurance Portability and Accountability Act authorization by site Institutional Review Boards. This was a retrospective study of lumbar spine imaging reports sampled from a prospectively assembled cohort of adults studying the effect of report content on subsequent treatment decisions (13). The cohort consisted of patients enrolled between October 2013 and September 2016 from four integrated health systems (Kaiser Permanente of Washington, Kaiser Permanente of Northern California, Henry Ford Health System, and Mayo Clinic Health System). We assembled a reference-standard dataset from the N = 178,333 index reports available on April 2016, randomly sampled stratified by site and imaging modality to obtain an approximately balanced sample of N = 413 x-ray reports and N = 458 MR reports (Table 1). Sample sizes were based on recommendations for NLP classification tasks (14) and also additionally justified by post hoc power calculations.

Our team had two neuroradiologists, a physiatrist, and a physical therapist, all with clinical expertise in spine disorders (JGJ, HTH, PS, SDR). This group jointly identified 26 imaging findings that are potentially related to LBP, including eight commonly found on radiology reports, and six considered clinically important (Table 2). We based the findings identified in this study on prior research, and these findings represent a wide range of radiological findings related to LBP (2,15,16). For each finding, the experts listed associated keywords, which are synonyms of the finding, and supplemented them with online searches of medical databases (17,18). The keyword list was expanded iteratively, where additional keywords were identified through the annotation process.

Our clinicians annotated each report for the presence or absence of each finding using a secure online interface developed with Research Electronic Data Capture (Appendix Figure A1) (19). Interpretations of findings were guided by how reports are likely understood by receiving physicians (20). To achieve consistency in interpretation, all four experts annotated 71 reports for the 26 findings. Then, the group reviewed annotations and discussed discrepancies until consensus was reached. We arrived at our initial 71 reports through an iterative training process, selecting reports based on report length, where sufficient consensus among readers was achieved after having reviewed 71 reports. The remaining 800 reports were each annotated by two separate experts. We created all possible pairs from four experts for a total of six expert pairs; each rater pair annotated about 133 reports. Expert pairs discussed and corrected any discrepancies in their annotations. As needed, the senior neuroradiologist (JGJ) provided final adjudications. As a post-annotation check, all report ratings were updated to reflect changes in the keyword list. The

TABLE 1. Reference-standard Annotated Dataset

	N in Dataset	Average Document Length (number of words)	Average Age in years	Men (%)
Kaiser Permanente of Washington				
x-ray	102	1073 ± 212	70.4 ± 13.9	40.2
MR	115	2141 ± 731	58.9 ± 14.4	51.3
Total	217	1639 ± 767	64.3 ± 15.2	46.1
Kaiser Permanente of Northern California				
x-ray	104	1054 ± 235	67.5 ± 16.8	39.4
MR	114	1953 ± 647	57.1 ± 15.0	47.4
Total	218	1524 ± 668	62.1 ± 16.9	43.6
Henry Ford Health System				
x-ray	103	1129 ± 323	67.2 ± 16.0	28.2
MR	115	2130 ± 986	59.2 ± 16.0	49.6
Total	218	1657 ± 901	63.0 ± 16.4	39.4
Mayo Clinic Health System				
x-ray	103	1094 ± 280	69.4 ± 16.2	38.8
MR	115	1722 ± 645	55.1 ± 15.4	41.7
Total	218	1425 ± 595	61.9 ± 17.3	40.4
All				
x-ray	413	1088 ± 266	68.6 ± 15.8	36.7
MR	458	1987 ± 782	57.6 ± 15.2	47.5
Total	871	1561 ± 746	62.8 ± 16.4	42.4

MR, magnetic resonance imaging.
Values after ± are standard deviations.

reference-standard dataset contains annotations of all 26 findings based on relevant keywords. For the doubly annotated 800 reports, we measured inter-rater agreement in the reference-standard dataset using Cohen's kappa (21).

NLP System

We implemented our NLP system with feature extraction, model development, and model validation steps (Fig 1). Our NLP system included routines for preprocessing steps such as section segmentation, sentence segmentation, and text normalization. The section segmentation code separated each report into history, body, and impression sections, but we excluded the history sections from analyses as we were primarily concerned with findings. The sentence segmentation routine split text into individual sentences based on sentence boundaries. The text normalization routine mapped spelling errors and variations into the same word stem, and reduced variation in modeling. To make formatting more consistent across radiology reports received from four separate health systems, two authors with expertise in text processing (WKT, ENM) conducted programmatic checks and data cleaning. We developed separate rule-based and machine-learned models for each of the 26 findings.

We implemented the rule-based model (Appendix Figure A2) in Java (v4.6.0), using Apache Lucene (v6.1.0), Porter Stemmer, and NegEx (22–25). For every sentence of each report, we searched for keywords using regular expressions, which are sequences of characters and symbols (26), to identify relevant terms (called *Regex*). Then, we used a publicly available algorithm (24) to infer whether keywords were negated (called *NegEx*). We included a rule favoring the impression section information over any conflicting information in the body section (27). We produced dichotomous predictions for each report, where a positive assignment was made if there was at least one sentence with a keyword that was not modified by a negation term.

We implemented the machine-learned model in R (v3.3.0) (28), using the caret (v6.0-73) package (29). Predictors used included sequences of N words (called *N-grams*), *section* tags, *Regex* and *NegEx* from the rule-based model, imaging *modality*, and health system *site* (Fig 2). Therefore, the machine-learned model used both input features based on the output of the rule-based model (*Regex* and *NegEx*), plus additional predictors. *N-grams* are considered baseline features in NLP applications; we additionally modeled whether the *N-grams* were in the body or impression section of reports. We included the categorical variables *modality* and *site* to account for any potential heterogeneity in language between x-ray and MR reports, as well as heterogeneity among different health systems. We excluded extremely rare (<0.05%) or common (>95%) *N-grams* and used logistic regression with elastic-net regularization as the modeling approach (30). Elastic-net is a variable selection procedure that penalizes both the absolute value and the squared magnitude of model coefficients, allowing simultaneous selection of correlated predictors. We randomly split the dataset into 80% development (training and validation) and 20% for testing to evaluate model diagnostic

TABLE 2. List of 26 Findings Identified by NLP System

Type of Finding	Imaging Finding	RadLex ID
Deformities	Listhesis—Grade 1*	RID4780‡
	Listhesis—Grade 2 or higher†	RID4780‡
	Scoliosis	RID4756
Fracture	Fracture	RID4658,4699,49608
	Spondylosis	RID5121
Anterior column degeneration	Annular Fissure*	RID4716-7, RID4721-3
	Disc Bulge*	RID5089
	Disc Degeneration*	RID5086
	Disc Desiccation*	RID5087
	Disc Extrusion†	RID5094-6
	Disc Height Loss*	RID5088
	Disc Herniation	RID5090
	Disc Protrusion*	RID5091-3
	Endplate Edema or Type 1 Modic†	RID5110
	Osteophyte—anterior column	RID5079
Posterior column degeneration	Any Stenosis	RID5028-34
	Facet Degeneration*	NA
Associated with leg pain	Central Stenosis†	RID5029-32
	Foraminal Stenosis	RID5034
	Nerve Root Contact	NA
	Nerve Root Displaced or Compressed†	NA
	Lateral Recess Stenosis†	NA
Nonspecific findings and other	Any Degeneration	RID5085
	Hemangioma	RID3969
	Spondylolysis	RID5120
	Any Osteophytes	RID5076,RID5078-9,RID5081

Any stenosis refers to stenosis at any location (central, foraminal, or lateral recess) or not otherwise specified. Any degeneration refers to any of disc degeneration, facet degeneration, or degeneration not otherwise specified.

* After a finding indicates the eight findings commonly found in subjects without LBP.

† Indicates the 6 findings that are less common but are potentially clinically important.

‡ Indicates the RadLexID for spondylolisthesis.

accuracy. We fine-tuned model hyperparameters on the development subsample, with 10-fold cross-validation using a receiver operating characteristic (ROC) loss function. We produced predicted probabilities, and applied 10-fold cross-validated thresholds to obtain dichotomous predictions.

Statistical Analyses

We conducted statistical analyses in R (v3.3.0). We measured the prevalence of each finding, as well as inter-rater agreements using Cohen's kappa (31), on the annotated dataset. We compared NLP model predictions to reference-standard annotations in the test set. Evaluation measures were sensitivity, specificity, F1-score and area under the ROC curve (AUC) (32). The F1-score is the harmonic mean (the reciprocal of the arithmetic mean of the reciprocals) of sensitivity and PPV, providing a single performance measure among reports with positive findings. For the rule-based models that outputted binary predictions, we used the trapezoidal rule approximation to the AUC (33). For the measures sensitivity, specificity, and AUC, we calculated point estimates for each finding as well as averaged overall findings, over the eight common findings, and over the six clinically important findings. We estimated 95% confidence intervals using bootstrap percentiles on the test set based on 500 iterations.

We sought to determine whether a rule-based or a machine-learned model was more accurate in classifying findings. Analyses focused on sensitivity and specificity. For each measure and for each finding, we compared the rule-based and machine-learned models using McNemar's test of marginal homogeneity, not accounting for multiple comparisons (34).

Among the machine-learned models, we also characterized model performance by evaluating both data availability (amount of data available to learn from) and linguistic complexity (difficulty posed by the learning problem). Analyses focused on the F1-score. Because the same development set was used for all findings, we used finding prevalence as a proxy for data availability. Because disagreements among experts were likely due to difficulty in identifying a finding from text, we used inter-rater agreement as a proxy for linguistic complexity. Due to the small number of findings ($N_{findings} = 26$), we did not provide inferential statistics; instead, we illustrated relationships with scatterplots and nonparametric local regression smoothed fits.

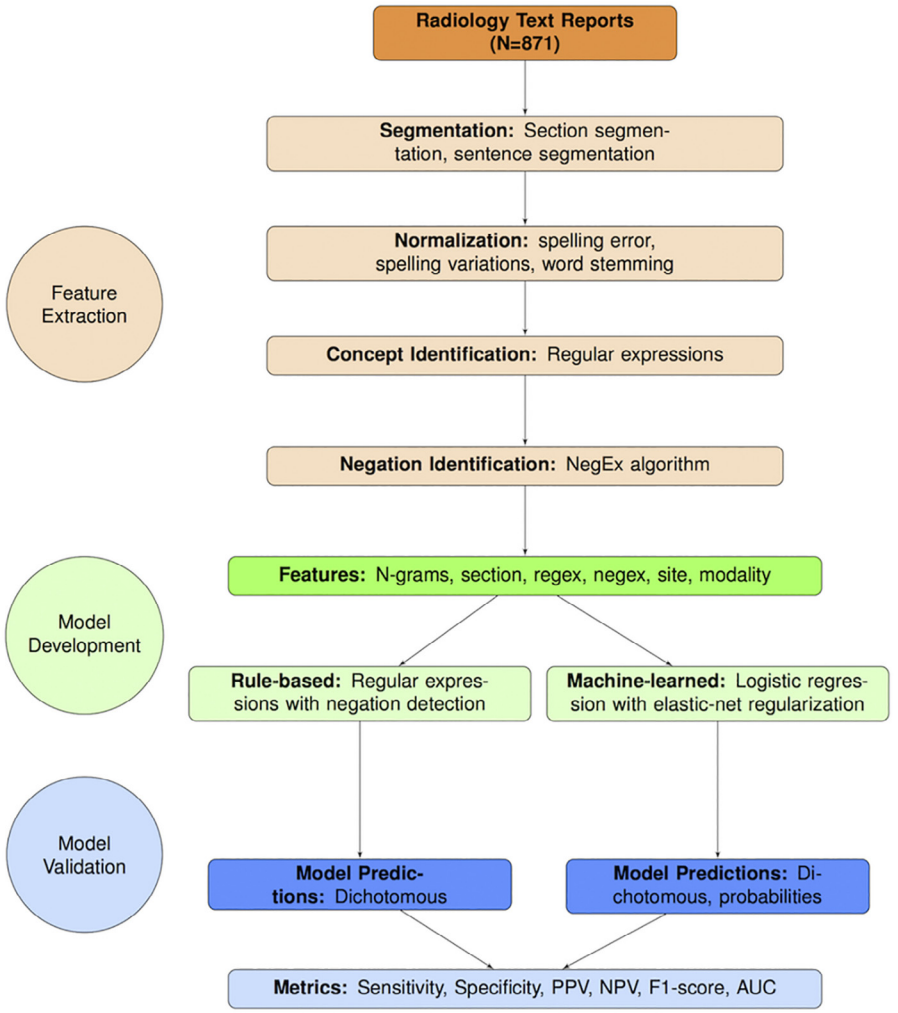

Figure 1. Flowchart illustrating steps involved in development of the natural language processing system on N = 871 medical expert annotated x-ray and magnetic resonance reports, sampled from four health system sites. Note: A "feature" is a natural language processing terminology that is equivalent to the terminology "predictor" in statistical modeling; "extraction" refers to the process of creating predictors from free text. AUC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value.

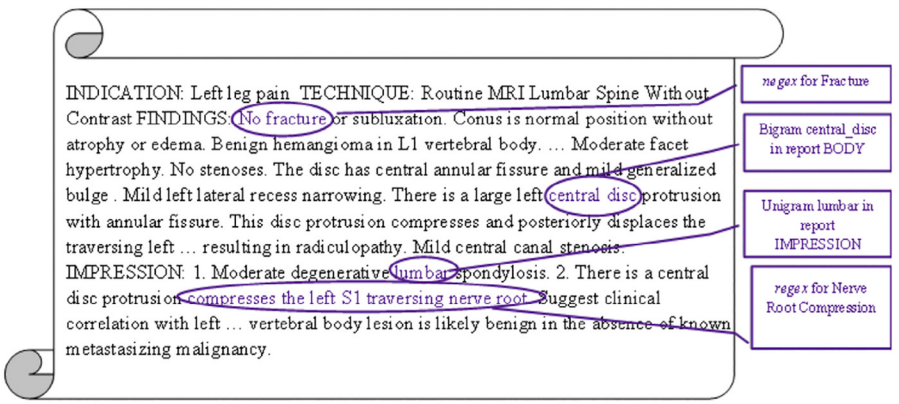

Figure 2. Examples of text-based predictors extracted from a radiology report snippet and used in machine-learned models. The phrase "no fracture" is used as a *NegEx* predictor (keyword negated) for the model to classify fracture. The phrase "compresses the left S1 traversing nerve root" is used as a *Regex* predictor (keyword present) for the model to classify nerve root displacement or compression. The N-grams "central disc" and "lumbar" are used as predictors for all machine-learned models.

Our study was Health Insurance Portability and Accountability Act compliant and approved by the Group Health Cooperative Institutional Review Board Protocol no. 476829 for the Lumbar Imaging with Reporting of Epidemiology (LIRE) pragmatic trial.

RESULTS

Appendices (Appendix A and Appendix B) are available online as supplementary material.

Reference-standard Dataset Characteristics

In the entire annotated dataset (N = 871), finding prevalence ranged from 3% (listhesis grade 2) to 89% (any degeneration). In the test set, finding prevalence ranged from 1% (listhesis grade 2) to 92% (any degeneration). Among the 800 reports, 25 of 26 findings achieved a minimum Cohen's kappa > 0.60, which is generally considered "substantial" agreement (31); we observed variation in agreement across findings

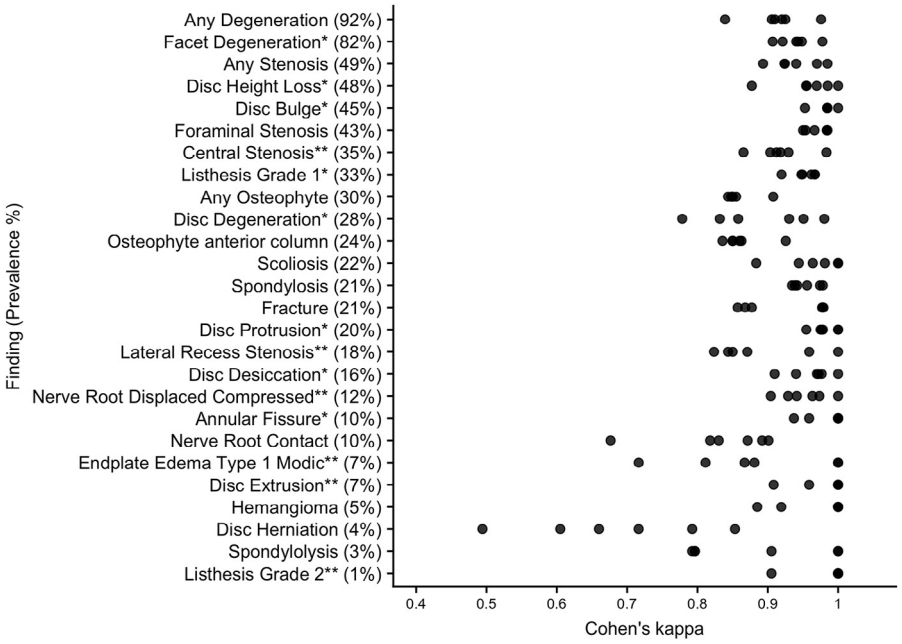

Figure 3. Distribution of agreement patterns in the annotated dataset. The findings are ordered by decreasing prevalence in the test set. Note: * after a finding indicates the eight findings commonly found in subjects without low back pain; ** indicates the 6 findings that are less common but are potentially clinically important.

as well as across rater pairs (Fig 3). We observed lower agreement for disc herniation (kappa = 0.49) and endplate edema (kappa = 0.72). Simple errors and report ambiguity accounted for most disagreements (Table 3).

NLP System Performances

For each of the 26 findings, Figure 4 illustrates patterns of model performance for sensitivity, specificity, and AUC; Figure 5 shows ROC curves for the eight findings that are common among patients without LBP; Appendix Table B1 summarizes the top five predictors from each machine-learned model.

Over all 26 findings in our study, the average sensitivity = 0.83 (95% confidence interval [CI] 0.38,1.00), specificity = 0.97 (95% CI 0.85,1.00), and AUC = 0.90 (95% CI 0.68,1.00) for the rules-based approach and average sensitivity = 0.94 (95% CI 0.76,1.00), specificity = 0.95 (95% CI 0.84,1.00), and AUC = 0.98 (95% CI 0.86,1.00) for the machine-learned approach. In detecting the eight findings commonly found in subjects without LBP, for the rules-based approach, average sensitivity = 0.90 (95% CI 0.71,1.00), specificity = 0.97 (95% CI 0.88,1.00), and AUC = 0.93 (95% CI 0.84,1.00) whereas for the machine-learned approach, average sensitivity = 0.95 (95% CI 0.82,1.00), specificity = 0.95 (95% CI 0.81,1.00), and AUC = 0.97 (95% CI 0.84,1.00). In detecting the six findings that are likely clinically more important for LBP, the rules-based average sensitivity = 0.57 (95% CI 0.00,1.00), specificity = 0.99 (95% CI 0.95,1.00), and AUC = 0.78 (95% CI 0.50,1.00) and the machine-learned average sensitivity = 0.96 (95% CI 0.86, 1.00), specificity = 0.96 (95% CI 0.85,1.00)., and AUC = 0.99 (95% CI 0.96,1.00). None of the machine-learned model had *site* among the top five predictors, but two findings (any stenosis, foraminal stenosis) included *modality* in the list of top 5 predictors (Appendix Table B1).

On average across all findings, the machine-learned approach provided improved sensitivity, comparable specificity, and an overall higher AUC. Among the rule-based models, the lowest sensitivities were observed for findings that consist of multiple concepts, for example lateral recess stenosis and nerve root compression or displacement. We provide more details on these points in the next subsection.

Comparison of Rules-based and Machine-Learned Models

Figure 6a shows the proportion of positive reports in the test set classified correctly by one model but not the other, compared to reference-standard true positives (sensitivity comparisons). Similarly, Figure 6b compares true negatives for each model (specificity comparisons); Appendix Table B2 illustrates 2 × 2 tables for calculating the proportions; Appendix Table B3 shows the resulting *P* values based on McNemar's tests.

For example, in Figure 6a, for the finding any stenosis (third row), among all of the reports classified as true positive, the rule-based and machine-learned models classified 19% differently. Of these 19%, the machine-learned model was correct for nearly all (94%) which is why the bar is nearly completely blue. In Figure 6b, among all the true negative reports, the rule-based and machine-learned models classified 11% of the reports differently, where the machine-learned model was correct for half. Overall, machine-learned models demonstrated substantial and statistically significant better sensitivity for 9 of 26 findings, and slightly worse model specificity for 7 of 26 findings that were statistically significant. Therefore, machine-learned models provided substantially higher

TABLE 3. Text Excerpts from Reference-Standard Dataset

Finding	Text Excerpts
Disc herniation	. . .degenerative change is evident at L2-L3 and. . . disc herniation is *not excluded*.
	Essentially unremarkable. L3-4: *Minimal* left posterior lateral focal herniation. . .
	right laminotomy. *No definite* disc herniation. Mild nonmasslike enhancing tissue. . .
Endplate edema or type 1 Modic	. . .S1 superior endplate with surrounding edema *suggesting* element of acuity. . .
	. . .high signal intensity on T2 and low signal intensity on T1 *suggestive of* acute to subacute superior endplate deformity.
	Minimal edema in the superior L5 endplate with more chronic appearance.
Lateral recess stenosis	Narrowing of the spine canal and lateral recesses and the right neuroforamen. . .
	. . .displaces the traversing left S1 nerve root in the left nerve root in the left lateral recess. . .
	. . .eccentric to the left with a left foraminal and far lateral component compressing the exiting left. . .
Nerve root displaced or compressed	Severe facet arthrosis with a diffusely bulging annulus causes moderate to severe central stenosis with redundant nerve roots above and below the interspace level.
	There is granulation tissue surrounding the descending right S1 nerve root. . .
	. . .has minimal mass effect on the descending left S1 nerve root. . .

Examples of report text from the reference-standard dataset show ambiguity in report text for the two findings with lower inter-rater agreement: Disc herniation (kappa = 0.49) and endplate edema (kappa = 0.72), and reports that were "missed" by rule-based but "found" by machine-learned models for lateral recess stenosis and nerve root displaced or compressed.
An ellipsis (. . .) indicates omitted raw text. Words in *italics* refer to ambiguous language.

sensitivities and slightly lower specificities compared to rule-based approaches.

The most substantial difference between the models though, was gains in using machine-learned models to detect the presence of compound findings. For example, nerve root displacement or compression can be thought of as a compound of the concepts "nerve root" and "displacement" or "compression," with varying language appearing between them in actual radiology reports. A rule-based model achieved a sensitivity of only 0.67, whereas a machine-learned model achieved sensitivity of 0.95 ($P < 0.05$). Similarly, the finding lateral recess stenosis, compound of the "lateral recess" and spinal "stenosis", was identified with a sensitivity of 0.41 by the rule-based model vs a sensitivity of 0.94 by the machine-learned model ($P < 0.05$). Table 3 illustrates examples of reports missed by rule-based but identified by machine-learned models for these two findings. In these cases, the compound parts of the findings can be inferred from context.

Effect of Data Availability and Linguistic Complexity on Machine-learned Models

Figure 7a displays the relationship between machine-learned model performance as measured by F1-score and data availability as measured by finding prevalence; Figure 7b shows model performance by linguistic complexity as measured by average Cohen's kappa.

Model performance was typically higher when more data were available for learning. Among the 13 findings with prevalence less than 20%, F1-score of machine-learned models was observed to increase substantially with the increase in finding prevalence. Among the 13 findings with prevalence greater than 20%, F1-score increased slightly with the increase in finding prevalence.

Model performance tended to be higher when the finding was less linguistically complex. Among the 17 findings with kappa greater than 0.85, F1-score of machine-learned models was observed to increase substantially with the increase in kappa (decrease in linguistic complexity). Among the nine findings with kappa less than 0.85, the effect of linguistic complexity on model performance was less definitive; however, the variability in F1-score was high.

DISCUSSION

Many imaging findings related to LBP are not explicitly coded in medical databases that are part of the electronic health record. NLP allows for automated identification of such findings from free-text radiology reports, reducing the burden of manual extraction (26). In this study, we sought to develop and validate an NLP system to identify 26 findings related to LBP from radiology reports. Our initial goal was to build a system for eight radiological findings common among subjects without LBP, and we subsequently expanded this goal to encompass additional relevant findings. Our reference-standard annotations were guided by how radiology reports are likely interpreted by receiving physicians, and our NLP system was developed accordingly (20). To ensure internal consistency, we conducted pre-annotation training, double annotation of reports, and post-annotation checking of labels, resulting in a dataset with high inter-rater reliability for most findings.

In practice, NLP approaches can include rule-based and machine-learned (26,32). Rule-based approaches have minimal set-up costs, but are burdensome to scale beyond limited findings or over time as language usage changes. Machine-learned approaches leverage the same feature sets to predict

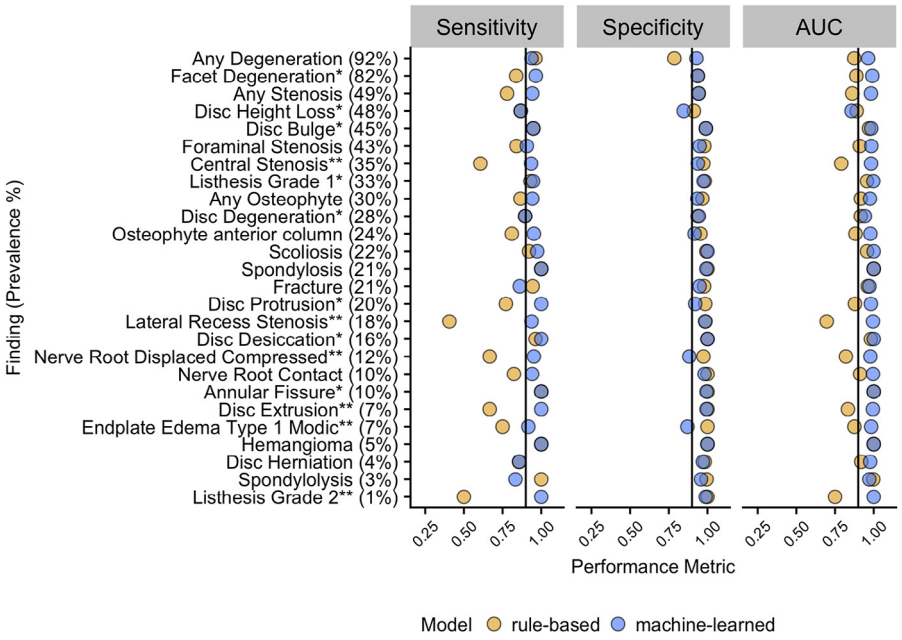

Figure 4. Point estimates of sensitivity, specificity, and AUC of rule-based and machine-learned models for each finding as measured in a test set of N = 174. The findings are ordered by decreasing prevalence in the test set; black lines on each panel correspond to 0.90. Note: * after a finding indicates the 8 findings commonly found in subjects without low back pain; ** indicates the 6 findings that are less common but are potentially clinically important. AUC, area under the receiver operating characteristic curve. (Color version of figure is available online.)

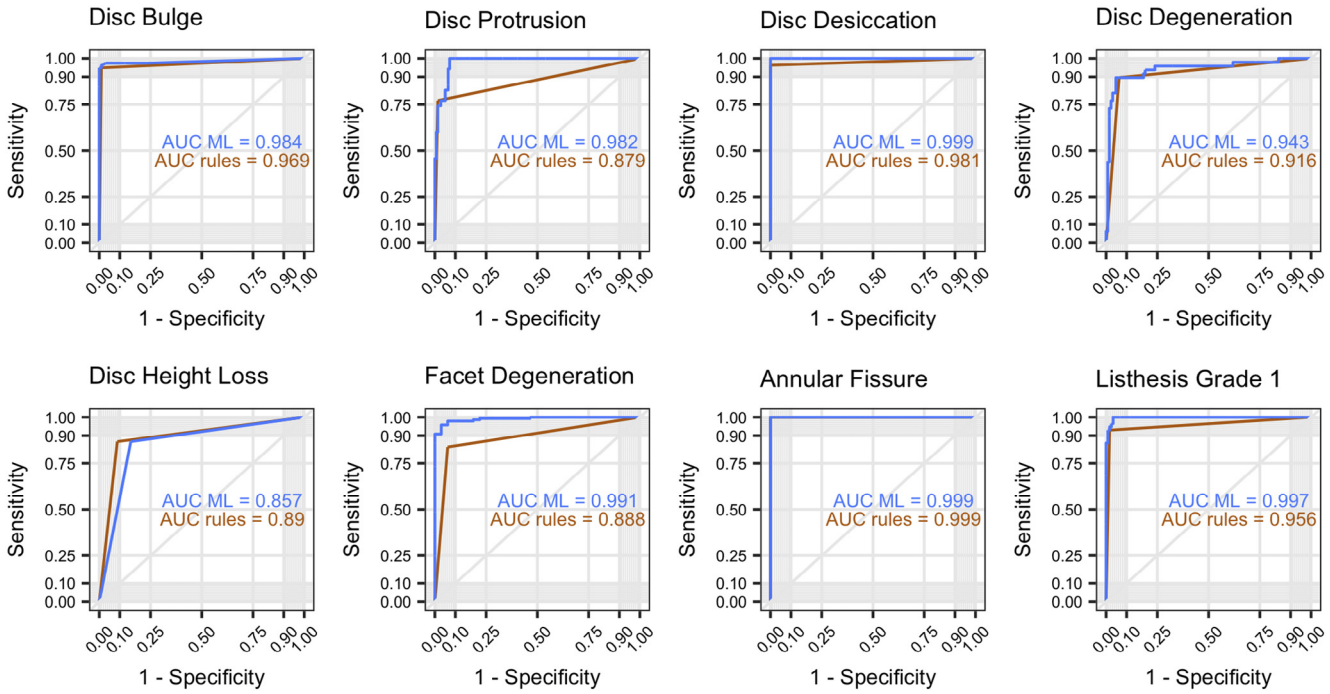

Figure 5. ROC curves and AUC in the test set for rule-based and machine-learned models of the eight findings commonly found in subjects without low back pain. AUC, area under the receiver operating characteristic curve; ML, machine-learned. (Color version of figure is available online.)

multiple findings, however require large sample sizes for development. The flexibility of machine-learned models may be more preferable to the rigidity of rules-based models, when considering scalability to large EMR databases. Furthermore, complicated machine-learned NLP models may not be necessary for report-level classification tasks: Zech et al reported that simple features such as *N-grams* together with logistic regression was as accurate as more sophisticated features (35).

Our rule-based models achieved moderate sensitivity and very high specificity, comparable to the performances of rule-based models to identify findings for other conditions (36). Such performance has been often attributed to usage of open-source negation detection algorithms (24) to reduce false positives. Our machine-learned models provided substantial gains in model sensitivity with slight loss in specificity, and had overall higher discrimination. Such gains are due to

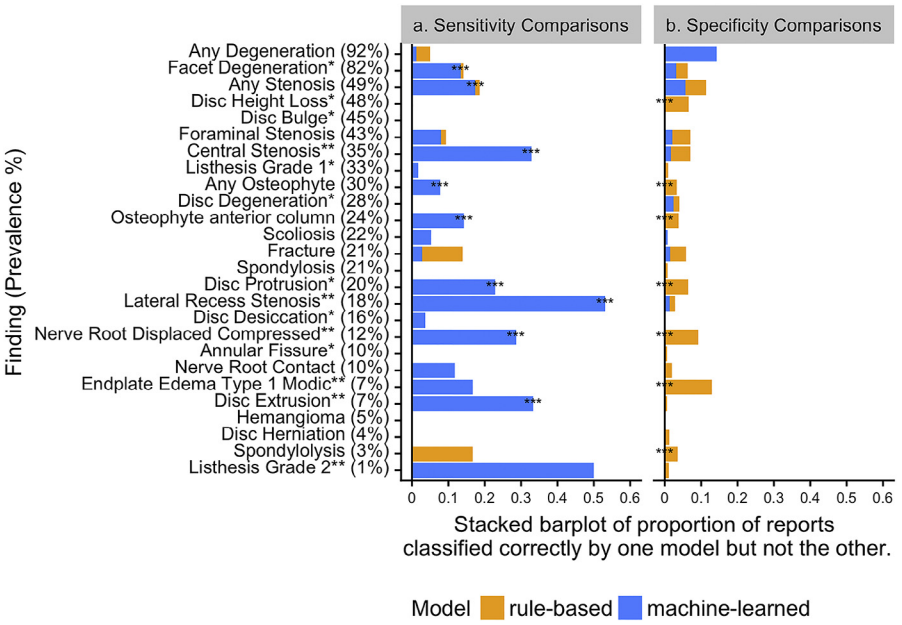

Figure 6. **(a)** Sensitivity comparisons: Proportion of reports classified correctly by one model but not the other, as compared to true positive annotations in the test set. **(b)** Specificity comparisons: Proportion of reports classified correctly by one model but not the other, as compared to true negative annotations in the test set. The test set has N = 174 reports; denominators in calculating proportions are test set size multiplied by prevalence (for sensitivity comparisons), or one minus prevalence (for specificity comparisons). Note: * after a finding indicates the eight findings commonly found in subjects without low back pain; ** indicates the six findings that are less common but are potentially clinically important. *** at the left of the stacked bar plots indicates $P < 0.05$ from McNemar's test, not adjusting for multiple comparisons. (Color version of figure is available online.)

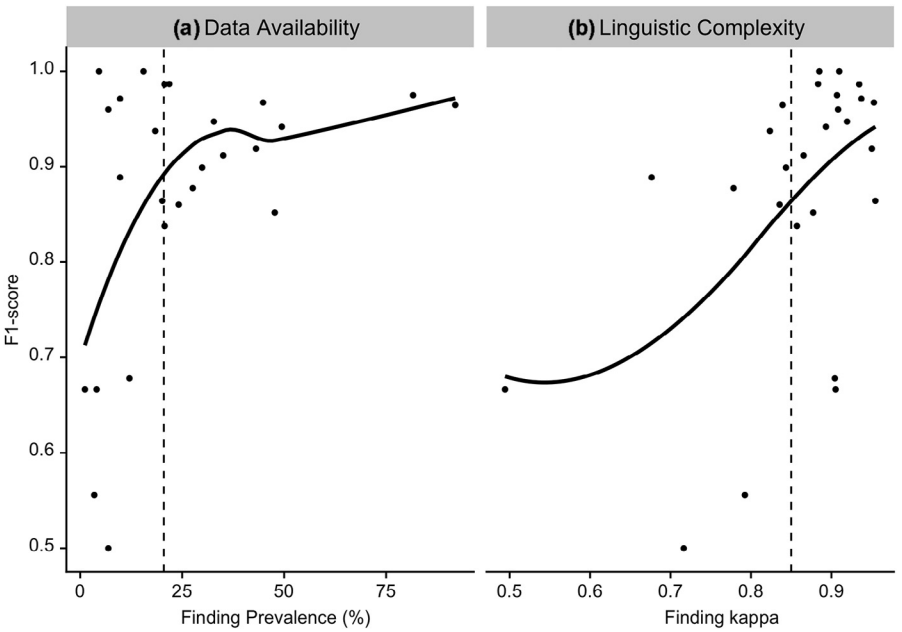

Figure 7. **(a)** Performance of machine-learned models as measured by F1-score vs data availability as measured by finding prevalence; vertical dashed line is finding prevalence = 20%. **(b)** Performance of machine-learned models as measured by F1-score vs linguistic complexity as measured by inter-rater kappa; vertical dashed line is kappa = 0.85. Note: Each dot represents a finding. The solid curves are nonparametric local regression fits to the data.

machine-learned models augmenting the predictions of rule-based models with additional contextual text-based predictors (Appendix B).

One concern in using NLP for biomedical applications is the generalizability to new study sites, because reporting style and structure could vary by institution and radiologist training (26). Our dataset was assembled across four health systems, and site was included as a predictor in the machine-learned models. Interestingly, no models had site in the list of top five predictors, indicating that language used in expressing the imaging findings was relatively consistent across study sites. However, within-site linguistic variability is possible, because each site in our study encompassed multiple clinics, radiologists, and data systems. Due to small sample sizes for each site (Table 1), we caution against any conclusive evidence.

Another issue in using NLP in radiology is the potential linguistic heterogeneity across imaging modalities. Our study included both x-ray and MR reports, where MR reports were on average twice as long as x-ray reports (Table 1). Additionally, although findings involving soft tissues and internal structures cannot directly be seen on x-ray images, they may occasionally be inferred and dictated on x-ray reports, especially when providing recommendations for symptomatic patients (27). Therefore, we annotated for all 26 findings on reports from both modalities. Our machine-learned models were developed on reports from both modalities due to sample

size constraints. We did, however, attempt to account for potential linguistic heterogeneity, by including modality as a predictor. Only two findings (any stenosis, foraminal stenosis) included modality in the list of top 5 predictors (Appendix Table B.1), suggesting that any linguistic heterogeneity across modalities were contained within report text, and that the additional information of modality did not improve prediction.

Our study included 26 findings of differing prevalence and linguistic complexity. We demonstrated that using the same development sample, performances of machine-learned models were poorer when a finding was less prevalent or linguistically more complex. These were instances when, due to the low number or quality of instances of a finding, there was insufficient information in the development data for machine-learned models to classify accurately (37). Our results suggest that even within the same NLP system, additional work required to fine-tune model performances could vary depending on characteristics of target findings.

A limitation of this study is that we required dichotomous annotations and predictions, even though radiology reports can contain ambiguous terms such as "suggesting" and "no definite" (Table 3) which may affect model performance (38). We attempted to reduce any potential bias by consistently coding such terms as indicating the presence of the finding. Another limitation is sample size. Even though our sample size was consistent with recommendations for classification tasks with moderate prevalence findings (14), the machine-learned models of rarer (prevalence <20%) findings had lower PPVs and were imprecise (Appendix Table A2), comparable to the results seen in Yetisgen-Yildiz et al (37). Larger datasets could provide larger training samples, thus better accuracies for machine-learned models, as well as larger testing samples, therefore validation performance measures that are less variable. We plan to augment our dataset with additional annotations in future work, using alternative sampling designs that can facilitate resource-efficient annotations. Another limitation is that we only considered x-ray and MR reports in this study, and thus our results may not generalize to other imaging modalities.

We recognize that our regularized regression-based machine-learned models were relatively simple, and deep learning neural networks could achieve higher accuracy if given a sufficiently large development sample size. In addition, instead of using NLP to identify lumbar spine imaging findings, image processing of actual scans has been observed to be highly accurate on MR images (39). It would be interesting to compare image processing and NLP predictions to downstream clinical outcomes.

Ultimately, our experience demonstrated the feasibility of an NLP system built with open-source tools to identify lumbar spine imaging findings from radiology reports sampled across two modalities and four health systems. We plan to use our NLP system to identify the eight findings that are common among subjects without LBP, which our system demonstrated high accuracy for in the testing sample, after further validation on additional reports sampled from the cohort.

CONCLUSIONS

We developed and validated an NLP system to identify 26 findings related to LBP from x-ray and MR radiology reports sampled from four health systems. Machine-learned models provided substantial increase in sensitivity with the slight loss of specificity compared to rule-based models. Model accuracies were affected by finding prevalence and finding complexity.

ACKNOWLEDGMENTS

This work is supported by the National Institutes of Health (NIH) Common Fund, through a cooperative agreement (5UH3AR06679) from the Office of Strategic Coordination within the Office of the NIH Director. The views presented here are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health. Dr. Suri is a staff physician at the VA Puget Sound Health Care System in Seattle, Washington. Dr. Suri is supported by VA Career Development Award #1IK2RX001515 from the US Department of Veterans Affairs Rehabilitation Research and Development Service. The contents of this work do not represent the views of the US Department of Veterans Affairs or the US Government.

REFERENCES

1. Hoy D, Bain C, Williams G, et al. A systematic review of the global prevalence of low back pain. Arthritis Rheumatol 2012; 64:2028–2037. https://doi.org/10.1002/art.34347.
2. Atlas SJ, Deyo RA. Evaluating and managing acute low back pain in the primary care setting. J Gen Intern Med 2001; 16:120–131. https://doi.org/10.1111/j.1525-1497.2001.91141.x.
3. Katz JN. Lumbar disc disorders and low-back pain: socioeconomic factors and consequences. J Bone Joint Surg 2006; 88(suppl 2):21–24. https://doi.org/10.2106/JBJS.E.01273.
4. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH Task Force on research standards for chronic low back pain. Eur Spine J 2014; 23:2028–2045. https://doi.org/10.1007/s00586-014-3540-3.
5. Costa Lda C, Maher CG, McAuley JH, et al. Prognosis for patients with chronic low back pain: inception cohort study. BMJ 2009; 339:b3829. https://doi.org/10.1136/bmj.b3829.
6. Johnsson KE, Rosen I, Uden A. The natural course of lumbar spinal stenosis. Clin Orthop Relat Res 1992; 279:82–86.
7. Henschke N, Maher CG, Refshauge KM, et al. Low back pain research priorities: a survey of primary care practitioners. BMC Fam Pract 2007; 8:40. https://doi.org/10.1186/1471-2296-8-40.
8. Hancock MJ, Maher CG, Laslett M, et al. Discussion paper: what happened to the 'bio' in the bio-psycho-social model of low back pain? Eur Spine J 2011; 20:2105–2110. https://doi.org/10.1007/s00586-011-1886-3.
9. Brinjikji W, Luetmer PH, Comstock B, et al. Systematic literature review of imaging features of spinal degeneration in asymptomatic populations. AJNR Am J Neuroradiol 2015; 36:811–816. https://doi.org/10.3174/ajnr.A4173.
10. Jensen TS, Karppinen J, Sorensen JS, et al. Vertebral endplate signal changes (Modic change): a systematic literature review of prevalence and association with non-specific low back pain. Eur Spine J 2008; 17:1407–1422. https://doi.org/10.1007/s00586-008-0770-2.
11. Danforth KN, Early MI, Ngan S, et al. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. J Thorac Oncol 2012; 7:1257–1262. https://doi.org/10.1097/JTO.0b013e31825bd9f5.

12. 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. https://doi.org/10.1148/radiol.2241011118.
13. Jarvik JG, Comstock BA, James KT, et al. Lumbar Imaging with Reporting of Epidemiology (LIRE)—protocol for a pragmatic cluster randomized trial. Contemp Clin Trials 2015; 45:157–163. https://doi.org/10.1016/j.cct.2015.10.003.
14. 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. https://doi.org/10.1148/radiol.2341040049.
15. Jarvik JJ, Hollingworth W, Heagerty P, et al. The Longitudinal Assessment of Imaging and Disability of the Back (LAIDBack) study: baseline data. Spine 2001; 26:1158–1166.
16. Birkmeyer NJ, Weinstein JN, Tosteson AN, et al. Design of the spine patient outcomes research trial (SPORT). Spine 2002; 27:1361.
17. Centers for Medicare & Medicaid Services. ICD-9-CM Diagnosis and Procedure Codes: Abbreviated and Full Code Titles . Cms.gov. [Online] Available at: https://www.cms.gov/medicare/coding/ICD9providerdiagnosticcodes/codes.html.
18. Radiology Society of North America. RadLex. RadLex. [Online] Available at: http://www.rsna.org/radlex.aspx.
19. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009; 42:377–381.
20. Kahn CE, Jr Langlotz CP, Burnside ES, et al. Toward best practices in radiology reporting. Radiology 2009; 252:852–856. Available at: https://doi.org/10.1148/radiol.2523081992.
21. Cohen JA. Educational and psychological measurement. In: Coefficient of agreement for nominal scales. Vol. 20. Thousand Oaks, CA: Sage Publications Sage CA, 1960; 37–46. http://dx.doi.org/10.1177/001316446002000104. ISSN.
22. Oracle Corporation. Java. Redwood Shores, CA: Oracle Corporation, 2016.
23. The Apache Software Foundation. Apache Lucene. Forest Hill, MD: The Apache Software Foundation, 2016.
24. Chapman WW, Bridewell W, Hanbury P, et al. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001; 34:301–310. https://doi.org/10.1006/jbin.2001.1029.
25. Porter MF. Porter Stemmer: The Porter Stemming algorithm Web site. [Online] January 2006. Available at: https://tartarus.org/martin/PorterStemmer/def.txt.
26. Cai T, Giannopoulos AA, Yu S, et al. Natural language processing technologies in radiology research and clinical applications. Radiographics 2016; 36:176–191. https://doi.org/10.1148/rg.2016150080.
27. Friedman PJ. Radiologic reporting: structure. Am J Roentgenol 1983; 140:171–172. https://doi.org/10.2214/ajr.140.1.171.
28. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2013.
29. Kuhn M. caret: Classification and Regression Training. R package version 6.0-73. https://CRAN.R-project.org/package=caret.
30. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005; 67:301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 159–174. https://doi.org/10.2307/2529310.
32. Pons E, Braun LM, Hunink MG, et al. Natural language processing in radiology: a systematic review. Radiology 2016; 279:329–343. https://doi.org/10.1148/radiol.16142770.
33. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol 1975; 12:387–415. https://doi.org/10.1016/0022-2496(75)90001-2.
34. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947; 12:153–157. https://doi.org/10.1007/BF02295996.
35. Zech J, Pain M, Titano J, et al. Reports, natural language–based machine learning models for the annotation of Clinical Radiology. Radiology 2018; 171093. https://doi.org/10.1148/radiol.2018171093.
36. Mendonça EA, Haas J, Shagina L, et al. Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform 2005; 38:314–321. https://doi.org/10.1016/j.jbi.2005.02.003.
37. Yetisgen-Yildiz M, Gunn ML, Xia F, et al. A text processing pipeline to extract recommendations from radiology reports. J Biomed Inform 2013; 46:354–362. https://doi.org/10.1016/j.jbi.2012.12.005.
38. Carrell DS, Halgrim S, Tran DT, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol 2014; 179:749–758. https://doi.org/10.1093/aje/kwt441.
39. Jamaludin A, Lootus M, Kadir T, et al. Automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. Eur Spine J 2017; 26:1374–1383. https://doi.org/10.1007/s00586-017-4956-3.

SUPPLEMENTARY DATA

Supplementary data related to this article can be found at https://doi.org/10.1016/j.acra.2018.03.008.