A Natural Language Processing System to Identify Lumbar Spine Imaging Findings Related to Low Back Pain from Radiology Reports

Wei Ling Katherine Tan[1], Saeed Hassanpour, PhD[2], Patrick Heagerty, PhD[1], Sean Rundell, DPT, PhD[1], Pradeep Suri, MD, MPH[3], Hannu Huhdanpaa, MD, MSc[4], Kathryn James, PA-C, MPH[1], David Carrell, PhD[5], Curtis Langlotz, MD, PhD[6], Nancy Organ[7], Eric Meier, MS[1], Karen Sherman, PhD, MPH[8], David Kallmes, MD[9], Patrick Luetmer, MD[9], Brent Griffith, MD[10], David Nerenz, PhD[11] and Jeffrey Jarvik, MD, MPH[1]

(1)University of Washington, Seattle, WA, (2)Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH, (3)University of Washington, Seattle, (4)Radia, Inc, Lynwood, (5)Group Health Cooperative, Seattle, WA, (6)Department of Radiology, Stanford University, Palo Alto, CA, (7)Center for Biomedical Statistics, University of Washington, Seattle, WA, (8)Kaiser Permanente Washington Health Research Institute, Seattle, WA, (9)Mayo Clinic, Rochester, MN, (10)Henry Ford Hospital, Detroit, MI, (11)Henry Ford Health System, Detroit, MI

Research Objective

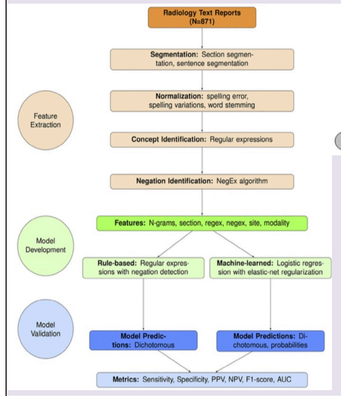
To evaluate a natural language processing (NLP) system built with open-source tools for identification of lumbar spine imaging findings in magnetic resonance (MR) and x-ray radiology reports.

Methods

Study Design and Population Studied

- Population studied: Lumbar Imaging with Reporting of Epidemiology (LIRE)[1] pragmatic randomized clinical trial, 4 US health systems.
  - Adult patients (≥ 18 y/o) whose primary care provider ordered x-ray or MR of the lumbar spine.
- Study design: Reference standard of N=871 radiology text reports dated between October 2013 and September 2016, stratified sampling by study site and imaging modality (x-ray or MR); each report annotated for the presence / absence of 26 findings.

Natural Language Processing (NLP)



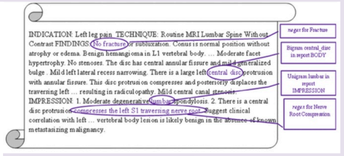


Figure 1. Flowchart illustrating steps involved in development of the NLP system on N = 871 medical expert annotated x-ray and magnetic resonance reports, sampled from four health system sites. Note: A "feature" is a NLP terminology that is equivalent to the terminology "predictor" in statistical modeling; "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.

Annotation: Double review by independent clinicians, adjudication by senior neuroradiologist

Machine-learning NLP: Training and evaluation on separate subsets (80%/20% of reference standard dataset)

Analysis

- Inter-rater agreement (Figure 3):
  - Data: Subset of reference standard (N=800).
  - Metric: Cohen's kappa for each annotator pair.
- NLP algorithm evaluation (Figure 4):
  - Data: 20% of reference-standard for testing (N=174).
  - Metrics: Sensitivity, Specificity, Area Under the Receiving Operating Characteristic (ROC) Curve (AUC) for each finding.
- Qualitative analyses (Table):
  - Radiology report Text excerpts with examples of ambiguous and complex language.

Results

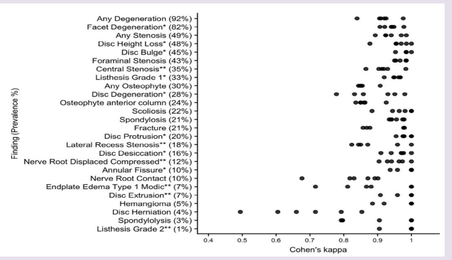


Figure 3. Distribution of agreement patterns in the annotated dataset. The findings are ordered by decreasing prevalence in the test set. Note: * after a finding indicates the eight findings commonly found in subjects without low back pain; ** indicates the six findings that are less common but are potentially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

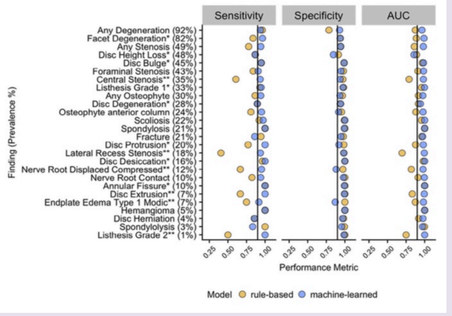


Figure 4. Point estimates of sensitivity, specificity, and AUC of rule-based and machine-learning models for each finding as measured in a test set of N = 174. The findings are ordered by decreasing prevalence in the test set; black lines on each panel correspond to 0.90. Note: * after a finding indicates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

Table. 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.

Limitations

- **Dichotomous NLP variables:** We required human annotation and NLP predictions to be binary (0 or 1), however radiology reports describe varying degrees of certainty.
- **Potential unaccounted heterogeneity:** We developed a single framework across imaging modalities, but there could be modality-specific differences, for example certain findings can only be seen on MR and not x-ray.
- **Clinical relevance:** Our NLP algorithm evaluation metrics are based on accuracy compared to reference-standard annotations. The clinical relevance of using such NLP predictions in practice depends on the research question.

Conclusions

- The described 26 radiological findings related to LBP have substantial agreement from medical experts, and accurately identified by NLP as benchmarked by reference-standard annotations
  - Machine-learned models provided substantial gains in model sensitivity with similar specificity, compared to rule-based models.
  - NLP algorithm accuracy is affected by ambiguous language and compound findings.
- Our results suggest that NLP algorithms and predictions can be integrated into large Electronic Medical Records (EMR) databases to identify patients with certain radiological findings related to LBP for clinical and research purposes.

References

[1]Jarvik, J.G., Comstock, B.A., James, K.T. et. al. 2015. Lumbar Imaging with Reporting of Epidemiology (LIRE)—protocol for a pragmatic cluster randomized trial. *Contemporary clinical trials*, *45*, pp.157-163.