

# **Systematic review of natural language processing (NLP) applications in magnetic resonance imaging (MRI)**

Gazal Mahmeed<sup>1</sup>, Dana Brin<sup>1</sup>, Eli Konen<sup>1</sup>, Girish N Nadkarni<sup>2,3</sup>, Eyal Klang<sup>2,3</sup>

1. Division of Diagnostic Imaging, Sheba Medical Center, Ramat Gan, Israel; Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel.

2. Division of Data-Driven and Digital Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

3. The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

## Corresponding Author:

Gazal Mahmeed, MD

Division of Diagnostic Imaging, Sheba Medical Center

Address: 2 Derech Sheba, Ramat Gan, Israel

Email: [Gazal.k.m@outlook.com](mailto:Gazal.k.m@outlook.com) ; Tel: +972-50-6982375

# Abstract

## *Background:*

As MRI use grows in medical diagnostics, applying NLP techniques could improve management of related text data. This review aims to explore how NLP can augment radiological evaluations in MRI.

## *Methods:*

We conducted a PubMed search for studies that applied NLP in the clinical analysis of MRI, including publications up to January 4, 2024. The quality and potential bias of the included studies were assessed using the QUADAS-2 tool.

## *Results:*

Twenty-six studies published between April 2010 and January 2024, covering more than 160k MRI reports were analyzed. Most of these studies demonstrated low to no risk of bias of the NLP. Neurology was the most frequently studied specialty, with twelve studies, followed by musculoskeletal (MSK) and body imaging. Applications of NLP included staging, quantification, and disease diagnosis. Notably, NLP showed high precision in tumor staging classification and structuring of free-text reports.

## *Conclusion:*

NLP shows promise in enhancing the utility of MRI. However, there is a need for prospective studies to further validate NLP algorithms in real-time clinical and operational scenarios and across various radiology specialties, which could lead to broader applications in healthcare.

## ***Introduction***

Natural language processing (NLP) combines computer science, artificial intelligence, and linguistics to improve how computers and humans interact. The introduction of technologies like ChatGPT in 2022 marked a major shift in NLP, showcasing its broad potential<sup>1</sup>.

In radiology, traditionally reliant on computer vision<sup>2, 3</sup> NLP introduces a new angle<sup>4, 5</sup>, with many potential uses, including flagging findings, prioritizing patients, generating imaging protocols, and conducting research<sup>6, 7</sup>.

MRI, known for its high contrast resolution and no radiation, is becoming more prevalent in diagnostic practices<sup>8</sup>. Using NLP in MRI interpretation could enhance workflows, diagnostics, and patient care.

This review assesses how NLP improves MRI applications by enhancing textual analysis in radiology.

## ***Methods***

This systematic review was reported according to the preferred reporting items for systematic reviews guidelines (PRISMA). The study is registered under PROSPERO, number (CRD42024518710).

### *Search strategy*

We searched literature to find studies on NLP's clinical uses in MRI. The search was conducted on January 4, 2024, using the PubMed database.

Search keywords included “MRI”, “Magnetic resonance imaging”, “MRE”, “Magnetic resonance enterography”, “NLP”, “Natural Language Processing”, “LLM”, “large language models”, and “chatGPT”. Details on complete search strategies are provided in (*Supplementary Material*).

Inclusion criteria were studies that (1) evaluated the clinical applications of NLP for MRI, (2) original articles in english (3) articles exclusively pertaining to MRI imaging.

We excluded (1) non - available full-text articles, (2) written in language other than english, (3) studies that included various imaging modalities other than MRI, (4) not focusing on NLP techniques, (5) studies focused on image processing and interpretation rather than text-based data analysis.

### *Study selection*

Two reviewers (GM, DB) independently screened the titles and abstracts to determine whether the studies met the inclusion criteria. The full-text article was reviewed when the title met the inclusion criteria or when there was any uncertainty. Disagreements were adjudicated by a third reviewer (EK).

### *Data extraction*

We used a standardized sheet to collect data on publication year, study design, location, database size, criteria, NLP methods, radiology field, MRI technique, NLP usage in MRI context, and performance.

### *Quality assessment and risk of bias*

Quality was assessed by the adapted version of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) criteria <sup>9</sup>. Details on quality assessment are provided in (*Supplementary Material*).

### *Data synthesis and analysis*

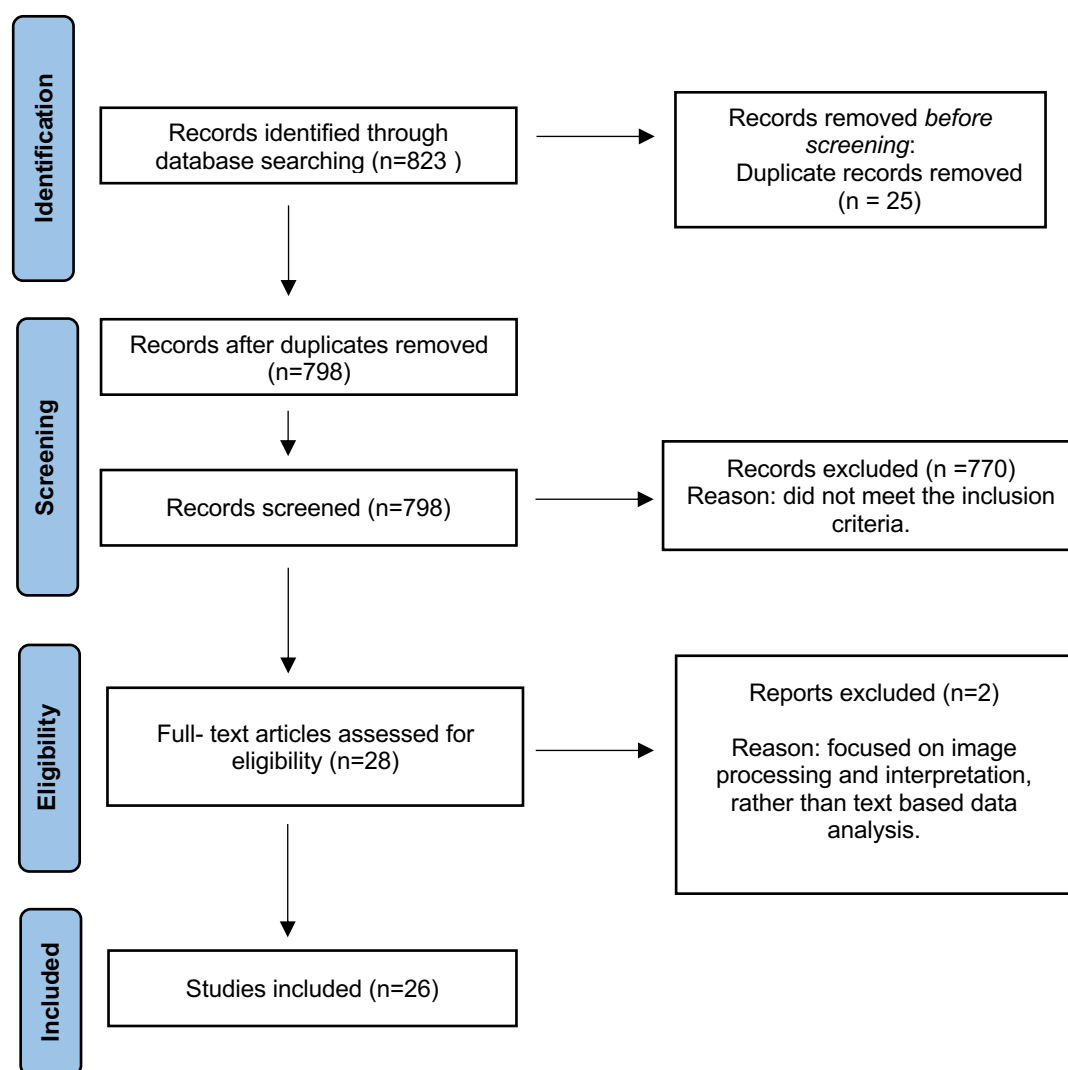
The analysis in this review is mainly qualitative. The heterogeneity of the studies in the literature evaluating NLP in MRI, their methods and the reported results precludes us from performing a meta-analysis.

## Results

### Study selection and characteristics

The initial search yielded 823 articles, with 26 meeting our inclusion criteria. **Figure 1** summarizes the characteristics of the included studies. The studies were published between 2010 and 2024. **Table 1** lists the publications reporting on the use of NLP in MRI.

**Figure 1:** PRISMA flowchart of the study selection process.



**Table 1:** Publications reporting on the use of NLP in MRI.

Ref	Title	Authors	Publication Year	Study Design	Radiology Field	Number of MRI Scans
<sup>10</sup>	Natural language processing for identification of hypertrophic cardiomyopathy (HCM) patients from cardiac magnetic resonance (CMR) reports	Dewaswala N, Arruda-Olson AM.	2022	Retrospective	Cardiac	391
<sup>11</sup>	Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke	Kim C, Lenert L.	2019	Retrospective case control	Neuro	3204
<sup>12</sup>	The implementation of natural language processing to extract index lesions from breast magnetic resonance imaging reports	Liu Y, Wang X.	2019	Retrospective	Breast	1633
<sup>13</sup>	Automated Radiology-Arthroscopy Correlation of Knee Meniscal Tears	Li MD, Huang AJ.	2022	Retrospective	MSK	3593

	Using Natural Language Processing Algorithms					
14	Using Natural Language Processing of Free-Text Radiology Reports to Identify Type 1 Modic Endplate Changes	Huhdanpaa HT, Jarvik JG.	2018	Retrospective	Neuro	458
15	Automatic Diagnosis Labeling of Cardiovascular MRI by Using Semi-supervised Natural Language Processing of Text Reports	Zaman S, Linton N.	2021	Retrospective	Cardiac	1503
16	A Scalable Natural Language Processing for Inferring BT-RADS Categorization from Unstructured Brain Magnetic Resonance Reports	Lee SJ, Banerjee I.	2020	Retrospective	Neuro	1519
17	Natural Language Processing for Automated Quantification of Brain Metastases Reported in Free-Text Radiology Reports	Senders JT, Arnaout O.	2019	Retrospective	Neuro	1479
18	Discerning tumor status from unstructured MRI reports completeness of information in existing reports and utility of automated natural language processing	Cheng LT, Erickson BJ.	2010	Retrospective	Neuro	778

19	Application of natural language processing to post-structuring of rectal cancer MRI reports	Liu W, Li Y.	2024	Retrospective	Body	358
20	Utility of a Rule-Based Algorithm in the Assessment of Standardized Reporting in PI-RADS	Zhang D, Gupta RT.	2023	Retrospective	Body	6264
21	Automatic extraction of imaging observation and assessment categories from breast magnetic resonance imaging reports with natural language processing	Liu Y, Wang XY.	2019	Retrospective	Breast	2330
22	Natural language processing deep learning models for the differential between high-grade gliomas and metastasis: what if the key is how we report them?	Martín-Noguerol T, Luna A.	2023	Retrospective	Neuro	185
23	Development and web deployment of an automated neuroradiology MRI protocoling tool with natural language processing	Chillakuru YR, Sohn JH.	2021	Retrospective	Neuro	8129
24	A Natural Language Processing-based Model to Automate MRI Brain Protocol Selection and Prioritization	Brown AD, Marotta TR.	2017	Retrospective	Neuro	13982
25	Automatic Determination of the Need for Intravenous Contrast in	Trivedi H, Sohn JH.	2018	Retrospective	MSK	1544



	Musculoskeletal MRI Examinations Using IBM Watson's Natural Language Processing Algorithm					
26	Automation of Protocoling Advanced MSK Examinations Using Natural Language Processing Techniques	Eghbali N, Ghassemi MM.	2023	Retrospective study	MSK	32372
27	Prediction of Stroke Outcome Using Natural Language Processing-Based Machine Learning of Radiology Report of Brain MRI	Heo TS, Kim C.	2020	Prospective	Neuro	2538
28	Deep-Learning-Based Natural Language Processing of Serial Free- Text Radiological Reports for Predicting Rectal Cancer Patient Survival	Kim S, Shin SJ.	2021	Retrospective	Body	4338
29	Machine learning outcome prediction using stress perfusion cardiac magnetic resonance reports and natural language processing of electronic health records	Alskaf E, Chiribiri A.	2024	Retrospective	Cardiac	6344
30	Natural language processing to predict isocitrate dehydrogenase genotype in diffuse glioma using MR radiology reports	Kim M, Sohn B.	2023	Retrospective	Neuro	1427

31	Effects of age and sex on the distribution and symmetry of lumbar spinal and neural foraminal stenosis: a natural language processing analysis of 43,255 lumbar MRI reports	Travis Caton M Jr, Andriole KP.	2021	Retrospective	Neuro	43255
32	A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports	Truhn D, Nebelung S.	2023	Retrospective	MSK	132
33	Comparing Artificial Intelligence Approaches to Retrieve Clinical Reports Documenting Implantable Devices Posing MRI Safety Risks	Valtchinov VI, Khorasani R.	2020	Retrospective	General	25,000
34	Developing a triage predictive model for access to a spinal surgeon using clinical variables and natural language processing of radiology reports	Krebs B, Gross DP.	2023	Historical cohort	Neuro	398
35	Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients	Chung EM, Kamrava M.	2023	Retrospective	Body	55

**Table 2:** Summary of Studies Applying NLP to MRI: Modalities, Methods, Models, Tasks, and Performance Scores.

Ref	NLP Clinical Application	MRI Modality	NLP Technique	NLP Task	Performance Score
<sup>10</sup>	Disease diagnosis	Cardiac	Rule based and machine learning	Data structuring	Accuracy 99%
<sup>11</sup>	Disease diagnosis	Brain	Comparison between multiple NLP models	Data structuring	Single decision tree showed the highest performance:  F1-measure 93.2%  Accuracy 98%
<sup>12</sup>	Disease diagnosis	Breast	Rule-based method	Data structuring	Identification of index lesion:  Recall and Precision > 85.0%
<sup>13</sup>	Disease diagnosis	Knee	Supervised machine learning models (logistic regression, SVM, and random forest)	Data structuring	Medial meniscus F1 scores 93%-94%  Lateral meniscus F1 scores 86%- 88%
<sup>14</sup>	Disease diagnosis	Spine	Rule-based NLP algorithm	Data structuring	Recall 70%  Specificity 99%  F1 79%

15	Disease diagnosis	Cardiac	Machine learning models	Data structuring	The BERT-based model achieved a micro-averaged F1 score 86%
16	Staging and quantification	Brain	Deep learning	Data structuring	Unstructured reports-f1 score 72%  structured reports-f1 score of 98%
17	Staging and quantification	Brain	Comparison between multiple NLP models	Data structuring	LASSO regression model demonstrated the best overall performance  AUC of 92%  Accuracy 83%  libration Intercept 6%
18	Staging and quantification	Brain	Statistical and rule-based methods	Data structuring	Sensitivity 80.6%  Specificity 91.6%
19	Staging and quantification	Rectum	Rule-based NLP model	Data structuring	Pre-2015 reports:  accuracy 93.8%  precision 95.6%  recall 87.1%  F1 score 91.2%

					<p>Post-2021 reports</p> <p>accuracy 92.5%</p> <p>precision 98.5%</p> <p>, recall 94.15%</p> <p>F1 score 96.3%</p>
20	Staging and quantification	Prostate	Rule-based NLP model	Data structuring	<p>Accuracy 92.6%</p> <p>Precision 88.8%</p> <p>Recall 85.6%</p> <p>F1 score 87%</p>
21	Staging and quantification	Breast	An internally developed NLP program	Data structuring	<p>Recall 78.5%</p> <p>Precision 86.1%</p>
22	Staging and quantification	Brain	Deep learning models CNN, BiLSTM, BERT	Data structuring	<p>CNN network provided the best results :</p> <p>Macro-avg precision 87.3%</p> <p>Sensitivity 87.5%</p> <p>F1 score 87.2%</p>
23	Protocol selection	Spine, Brain	machine learning models (FastText, XGBoost)	Data structuring	<p>Spine MRI model:</p> <p>Accuracy 83.4%</p> <p>AUC 88%.</p>

					<p>The head MRI :</p> <p>Accuracy 85.4%</p> <p>AUC 94%</p> <p>contrast brain protocol</p> <p>AUC 92%</p>
24	Protocol selection	Brain	Machine learning models (random forest, support vector machine (SVM), and k-nearest neighbor (KNN).)	MRI protocols	<p>Accuracy:</p> <p>Protocol selection 82.9%</p> <p>Contrast administration</p> <p>83.0%</p> <p>Prioritization tasks 88.2%</p>
25	Protocol selection	MSK, Spine	Comparison between multiple NLP models	MRI protocols	<p>Watson vs. original protocol Sensitivity 74%</p> <p>Accuracy 83%</p> <p>Watson vs. second reader Sensitivity 81%</p> <p>Accuracy 88%</p> <p>Watson vs. original and second reader agreed case only reader Sensitivity</p> <p>83%</p> <p>Accuracy 90%</p>

26	Protocol selection	MSK	Comparison between multiple NLP models	MRI protocols	Accuracy 83%  AUC 87%
27	Prognosis prediction	Brain	Machine learning models	Prediction models	RF algorithm had the best  AUC 78%
28	Prognosis prediction	Rectum	Deep learning	Prediction models	N/A
29	Prognosis prediction	Stress perfusion cardiac magnetic resonance	Machine learning models	Prediction models	support vector machine (SVM) was the best:  F1 score 24%  AUC 80%
30	Pathology prediction	Brain	Comparison between multiple deep learning models (LSTM , BiLSTM, BERT , BERT GCN , and BioBERT)	Prediction model	BERT GCN showed the highest performance:  AUC 85%-95%  CI 81% -89%
31	Comparative analysis	Lumbar spine	Rule-based natural language processing	Data structuring	Random sample of 100 LMRI reports  Accuracy 94.8%  At individual levels,  Accuracy ranged from 86.%

					at right L5-S1 to 100% in 5/18 level instances 27.8%.
32	Clinical decision support	Knee, Shoulder	Large language models (chatGPT)	Treatment recommendations	N/A
33	Safety protocol compliance	Multiple	Expert-driven NLP; ontology-driven NLP.	MRI protocols	MRI-Red:  - Ontology-Derived:  - Sensitivity: 96%  - Specificity: 90%  - Accuracy: 91%   MRI-Yellow:  - Ontology-Derived:  - Sensitivity: 76%  - Specificity: 62%  - Accuracy: 66%
34	Treatment recommendations	Spine	N/A	Prediction models	Nagelkerke R – squared  R2 = 20%
35	MRI report summary	Prostate	Large language models (chatGPT)	Data structuring	N/A



**Table 3: Summary of Main Results and Limitations of Synthesized Studies.**

Ref	Main Findings	Study Limitations
10	NLP identified and classified HCM from CMR narrative text reports with high performance.	<ol style="list-style-type: none"> <li>1. Complex sentences, ambiguity, and Sentences recorded in incorrect sections of the report were reasons for false-positive results.</li> <li>2. The NLP algorithms used were developed and tested in a single tertiary medical center in a cohort of patients with suspected HCM.</li> </ol>
11	Supervised ML based NLP algorithms are useful for automatic classification of brain MRI reports for identification of acute ischemic stroke (AIS) patients. Single decision tree was the best classifier to identify brain MRI reports with AIS.	<ol style="list-style-type: none"> <li>1. The text corpus was created at a single institution.</li> <li>2. Only included brain MRI reports with conventional stroke MRI sequence. In clinical practice, full conventional brain MRI sequence could vary depending on the degree of emergency in each situation, the patient's condition, and the laboratory results.</li> <li>3. The performance of machine learning (ML) classifiers could be affected by the class proportions in the training dataset. The proportion of brain MRI reporting in AIS may vary significantly depending on the characteristics of each hospital.</li> </ol>
12	NLP method successfully extracted the index lesion and its corresponding information from free-form text from breast MRI.	<ol style="list-style-type: none"> <li>1. The NLP system performed was not suitable for other reports that did not use BI-RADS descriptors.</li> <li>2. Only 7 out of 478 cases reported no lesion in the data, and there is no case with BI-RADS 0 or 1 in the rest of</li> </ol>

		<p>471 reports which reflected the general population who underwent breast MRI.</p> <p>3. The NLP system extracted the index lesion based on the authors' hypothesis that the index lesion accounts for the largest number of imaging features. This hypothesis was artificially defined and was not the direct extraction of the index lesion.</p>
13	<p>Radiology-arthroscopy correlation can be automated for knee meniscal tears using NLP algorithms, which shows promise for education and quality improvement.</p>	<p>1. Some studies without a disagreement between the knee MRI and arthroscopy reports would not be screened out by the NLP approach, which would then still require manual review.</p> <p>2. Data from a single institution.</p> <p>3. In analyzing MRI and arthroscopy correlation, there is a delay between the two studies and patients may develop new meniscal abnormalities in that time interval.</p> <p>4. Different types of meniscal tears may have different management implications</p> <p>5. The focus of the study is on meniscal tears, but other abnormalities may be clinically relevant for radiology-arthroscopy correlation and would require further labeling of training data.</p>
14	<p>Rule-based NLP is efficient approach for identifying patients with Type 1 Modic change if the emphasis is on identifying only relevant cases with low concern regarding false negatives</p>	<p>1. Difficulty of eliciting all possible keywords given the enormous variability of how lumbar spine findings are reported.</p> <p>2. Limited number of reports for a specific finding, findings that are complex to identify, ambiguity in reports, and feature sets which are not sufficiently rich.</p>

		<ol style="list-style-type: none"> <li>F1 scores are measures that are limited to being a relative term with no absolute range or ranges of poor, fair, good, or excellent.</li> <li>The developing rule-based algorithms is determining how far apart key words can be in a sentence to be considered together.</li> </ol>
15	The developed model used labels extracted from radiology reports to provide automated diagnosis categorization of cardiac MR images with a high level of performance.	<ol style="list-style-type: none"> <li>The BERT model was not 100% accurate for all five diagnoses.</li> <li>Researchers can use their model.</li> <li>They did not explore alternative ways to overcome the 512-token input limitation of BERT.</li> </ol>
16	Proposed NLP pipeline is capable of inferring BT-RADS report scores from unstructured reports after training on structured report data. The study provides a detailed experimentation process and may provide guidance for the development of RADS-focused information extraction (IE) applications from structured and unstructured radiology reports.	<ol style="list-style-type: none"> <li>It is a single-center, retrospective study.</li> <li>The grouping of BT3 subcategories raises another limitation of the study, as the distinction between them is clinically important, differentiating between pseudo-progression and likely true tumor progression.</li> <li>Some Model performed on unstructured reports.</li> </ol>
17	Among various NLP techniques, the bag-of-words approach combined with a LASSO regression model demonstrated the best overall performance in extracting binary outcomes from free-text clinical reports. This study provides a framework for the development of machine learning-based	<ol style="list-style-type: none"> <li>A consensus in human classification was used as ground truth, which is a commonly used method to generate an approximation in the absence of actual ground truth.</li> <li>Complete data set was manually classified to generate labels for training and testing, however when an NLP model will be used, only a minor portion will be labeled manually to predict the labels on the remaining data set.</li> <li>Models trained on data from single institutions.</li> </ol>

	NLP models as well as a clinical vignette of patients diagnosed with brain metastases.	
18	NLP demonstrated good accuracy for tumor status classification and may have novel application for automated disease status classification from electronic databases.	<ol style="list-style-type: none"> <li>1. limitations exist for unbalanced datasets where class sizes differ significantly.</li> <li>2. As imaging features and terminology vary between different tumor types and imaging modalities, the algorithm may not be applicable to different patient populations.</li> <li>3. The lack of uniformity across reports made complete alignment impossible.</li> </ol>
19	The NLP system with rule-based pattern matching achieved rapid and accurate structured processing of rectal cancer MRI reports. MRI reports with structured templates are more suitable for NLP-based extraction of information.	<ol style="list-style-type: none"> <li>1. The dataset came from a single center.</li> <li>2. MRI reports were based on the reading radiologists' interpretations, and the accuracy of the test dataset can be expected among the radiologists.</li> <li>3. The low reporting rate of all image features in reports before 2015 may have resulted in a lack of representativeness.</li> </ol>
20	Rule-based processing is an accurate method for the large-scale, automated extraction of PI-RADS scores from the text of radiology reports. These natural language processing approaches can be used for future initiatives in quality improvement in prostate mpMRI reporting with PI-RADS.	<ol style="list-style-type: none"> <li>1. RegEx algorithm was developed based on the reporting characteristics of a single institution.</li> <li>2. RegEx algorithm was unable to categorize prostate mpMRI reports in 5.49% of cases.</li> <li>3. The algorithm is unable to assess the underlying accuracy of the PI-RADS score assigned during the original clinical interpretation.</li> </ol>
21	The NLP algorithm demonstrates high recall and precision for information	<ol style="list-style-type: none"> <li>1. A large amount of the breast MRI reports was selected from their department and thus the segmentation rules</li> </ol>

	<p>extraction from free-text reports. This approach will help to narrow the gap between unstructured report text and structured data, which is needed in decision support and other applications.</p>	<p>and the reviewed BI-RADS lexicon were developed based on the writing habits of single department.</p> <ol style="list-style-type: none"> <li>There are often multiple lesions in breast MRI reports and the index lesion is most crucial to clinicians in determining the management and prognosis of patients. However, this study extracted information from all the lesions in breast MRI reports not just from the index lesions.</li> </ol>
22	<p>A deep learning model based on CNN enables radiologists to discriminate between high – grade glioma (HGG) and metastasis based on MRI reports with high-precision values.</p>	<ol style="list-style-type: none"> <li>insufficient number of radiology reports selected for the training and testing of the NLP tools .</li> <li>the translation from Spanish to English reports would have some kind of impact on the outcome of language of NLP tool as linguistic nuances are probably being missed during the translation process.</li> <li>lack of additional lesions on the radiology other than glioma.</li> </ol>
23	<p>The two NLP models developed accurately predict spine and head MRI protocol assignment, which could improve radiology workflow efficiency.</p>	<ol style="list-style-type: none"> <li>NLP models determine protocol assignment by word and word-context relationships. which can lead to unintended use of non-medically relevant, human biases hidden in the data.</li> <li>The head MRI protocol data lacked sufficient sample size on more specialized protocols.</li> <li>Data comes from a single academic institution.</li> </ol>
24	<p>NLP models developed from the narrative clinical information provided by referring clinicians and demographic data are feasible</p>	<ol style="list-style-type: none"> <li>Data comes from a single academic institution.</li> <li>For most of the study period, requisitions were completed by hand and the study relies on the faithful reproduction</li> </ol>

	<p>methods to predict the protocol and priority of MRI brain examinations.</p>	<p>of handwritten text by administrative staff or the interpreting radiologist</p> <p>3. Absence of important inputs like patient allergy information, glomerular filtration rate, or diabetes status, must rely on the referring service and age of the patient to predict the administration of gadolinium.</p> <p>4. Although protocol guidelines exist at their institution, it is not possible to completely remove variation in protocol selection.</p> <p>5. Due to the limitations of the picture archiving and communication system, a smaller subset of the dataset was used to evaluate the priority model.</p>
25	<p>A natural language classification algorithm can be trained with IBM Watson to automatically determine the need for intravenous contrast in musculoskeletal MRIs.</p>	<p>1. Challenges with spelling, grammar, and ambiguity in clinical indications.</p> <p>2. Difficulty troubleshooting errors due to a "black-box" algorithm.</p> <p>3. Issues with contrast assignment, potentially leading to clinical consequences.</p> <p>4. Lack of consideration for the requested study type, such as "MRI lumbar spine without contrast."</p> <p>5. Constraints in accessing and modifying IBM Watson's closed cloud service.</p> <p>6. Risks of harmful errors from potential updates to the service's algorithm.</p> <p>7. An intrinsic limitation in the scalability of our methods at the institution was the assignment of MRI protocols (which serves as ground truth) as free-text.</p>

26	<p>Application of NLP-based techniques has the potential to significantly reduce the time and cost spent on protocoling appropriate examinations. The results indicate that the proposed model can automate the assignment of orders for further revision. The low number of false negative instances suggests the reliability of the model</p>	<ol style="list-style-type: none"> <li>1. Limited number of protocols were implemented.</li> <li>2. Limitation in the generalizability of the model.</li> <li>3. The cost analysis is solely based on economic feasibility that is not considering the costs associated with potential patient harm and delay in treatment due to a false negative.</li> </ol>
27	<p>NLP-based deep learning (DL) algorithms can be used as an important digital marker for unstructured electronic health record data DL prediction.</p>	<ol style="list-style-type: none"> <li>1. MRI text report was read by neuroradiologists in single institution.</li> <li>2. They couldn't conclude whether these DL algorithms will perform better in predicting poor outcomes using brain MRI text reports in languages other than English.</li> </ol>
28	<p>The deep-transfer-learning model using free-text radiological reports can predict the survival of patients with rectal cancer, thereby increasing the utility of unstructured medical big data.</p>	<p>It was a retrospective study conducted at a single institution.</p>
29	<p>AI-based data extraction and analysis tools are becoming more available and bringing new insights into unstructured health data records. This study shows that this approach is feasible and reveals plausible data. With this they were able to confirm the predictive value of SP-CMR for CAD assessment. Positive ischaemic test and the presence of late gadolinium enhancement (LGE) was</p>	<ol style="list-style-type: none"> <li>1. A significant number of cases were excluded thus it would be expected to inflate the importance of CMR measures and produces some bias of the results.</li> <li>2. The NLP model in this study was not exclusive and some data extraction was performed using CogStack search engine of instances in texts and structured fields.</li> </ol>

	<p>associated with a higher risk of mortality regardless of clinical risk factors.</p>	<ol style="list-style-type: none"> <li>CMR reports at the time of this study were not included in NLP training and manual extraction was performed.</li> <li>Single large center study and only tested on one population.</li> </ol>
30	<p>bidirectional encoder representations from transformers, graph convolutional network [BERT][GCN] was externally validated to predict isocitrate dehydrogenase (IDH) mutation status in patients with diffuse glioma using routine MR radiology reports with superior or at least comparable performance to human reader.</p>	<ol style="list-style-type: none"> <li>There was a difference in the proportion of patients with IDH mutation between the two hospitals used for model development and patients with IDH mutation in the external validation higher proportion of set. This distribution shift may hamper fair performance evaluation of the models.</li> <li>The radiology reports since the introduction of electronic health record at the respective hospitals were included dating back to 2009 when the relevance and prognostic implication of IDH mutation were just being reported IDH mutation was incorporated in WHO classification of central nervous system tumor in 2016 with increasing efforts to associate imaging findings to molecular makers.</li> <li>sample size calculation was not performed when designing this study.</li> <li>While long short-term memory (LSTM) enables estimation of relative importance of text variables, it still remains largely unknown the based models extract and synthesize -exact mechanism transformer information on free texts.</li> </ol>
31	<p>NLP can identify patterns of lumbar spine degeneration through analysis of a large corpus of radiologist interpretations. Demographic differences in stenosis</p>	<ol style="list-style-type: none"> <li>The NLP algorithm empowers a much larger corpus to be analyzed, but its classification is imperfect.</li> </ol>



	<p>prevalence shed light on the natural history and pathogenesis of Lumbar spine degenerative disease( LSDD).</p>	<ol style="list-style-type: none"> <li>2. Their rules-based algorithm relied on a manually assembled dictionary mapping non-standard terminology for stenosis grading to A 6-point standard scale. This methodology is imperfect and may affect our estimate of the severity distribution of LSDD.</li> <li>3. The study cohort is not a true sampling of the general population as only patients with symptoms substantial enough to warrant imaging are studied.</li> </ol>
32	<p>GPT-4 demonstrates the potential to provide largely accurate and clinically useful treatment recommendations for common orthopedic knee and shoulder conditions. Expert surgeons rated the recommendations at least as "good", but the patient's situation and treatment urgency were not fully considered. Therefore, patients need to consult healthcare professionals for personalized treatment recommendations, while GPT -4 may be a supplementary resource rather than a replacement for professional medical advice after regulatory approval.</p>	<ol style="list-style-type: none"> <li>1. They studied only a few patients.</li> <li>2. To enhance its depth and relevance to clinical scenarios, GPT-4's predictions need to be more specific.</li> <li>3. The patient spectrum was broad.</li> <li>4. Treatment recommendations were qualitatively judged by two experienced orthopedic surgeons, and involvement of more surgeons could have strengthened the outcome basis even further.</li> <li>5. The tendency of GPT-4 to give generic and unspecific answers and to err on the side of caution rendered it challenging to assess its adherence to guidelines or best practices exactly.</li> <li>6. They used a standardized and straightforward way of prompting GPT-4. After more extensive modifications of these prompts, outcomes may be different.</li> </ol>
33	<p>Artificial intelligence approaches such as expert-driven NLP and ontology-driven NLP have similar accuracy in identifying</p>	<ol style="list-style-type: none"> <li>1. Unavailability of actual clinical outcomes for radiology department with evaluation.</li> <li>2. The study was limited to clinical notes obtained from single academic medical center.</li> </ol>

	patients with implantable devices that pose high safety risks for MRI.	
34	A logistic regression model was created to predict which patients may require spine surgery. Simple clinical variables appeared more predictive than variables created using NLP.	<ol style="list-style-type: none"> <li>1. A low sample size for this type of analysis.</li> <li>2. Data came from one spine assessment clinic, and the model has not been externally validated.</li> <li>3. By design, it relies on retrospective data, which can be prone to misclassification bias.</li> <li>4. The surgical outcomes of the patients in the study are unknown, thus it's not clear whether the surgical decisions made were optimal.</li> </ol>
35	Application of ChatGPT to summarize MRI reports at a reading level appropriate for patients. Physicians were likely to be satisfied with the summarized reports with respect to factual correctness, ease of understanding, and completeness. Physicians were less likely to be satisfied with respect to potential for harm, overall quality, and likelihood to send to patients.	<ol style="list-style-type: none"> <li>1. Small sample size and a single institution study.</li> <li>2. They focused exclusively on prostate cancer MRI reports, and the results may not be directly applicable to other types of radiology reports or cancer diagnoses.</li> <li>3. The anonymous questionnaire used in the study may be subject to response bias, as physicians with strong opinions on AI-generated summaries may be more likely to participate.</li> <li>4. There were concerns regarding "potential for harm, overall quality, and likelihood of sending the report to patients. The study is unable to identify potential causes for this discrepancy, and it is beyond the current scope of the study.</li> <li>5. It's possible that physicians communicated results with patients and did not document this in the electronic medical record.</li> </ol>

## *Descriptive summary of results*

Over the past years there has been an increase in the number of publications of NLP applied to MRI.

**Figure 2** represents this general trend, with a peak of publications in 2023. All studies included were retrospective in nature, encompassing 163,209 MRI reports.

**Figure 3** presents the distribution of radiology fields in the study. Neurology was the most prominent field, accounting for 12 out of 26 of the reviewed studies <sup>11 14 16 17 18 22 23 24 27 30 31 34</sup>. Musculoskeletal <sup>13 25 26 32</sup> and body <sup>19 20 28 35</sup> imaging were studied in four publications each, making them the second most represented fields. Cardiac imaging followed with three research studies <sup>10 15 29</sup>, while two studies involved breast imaging <sup>12 21</sup>. Additionally, there was one study that discussed the use of NLP for MRI in a more general context, without focusing on a specific field <sup>33</sup>.

The studies included various NLP techniques, such as rule based approaches, machine learning, and deep learning, including large language models (LLMs). The complete list of NLP techniques is detailed in **Table 2**.

We identified numerous valuable clinical applications of NLP in MRI. These applications are summarized in **Figure 4**, The key findings from each study are extensively described in **Table 3**.

NLP was found to have diverse applications in research studies, predominantly in staging, quantification, disease diagnosis, and protocol selection.

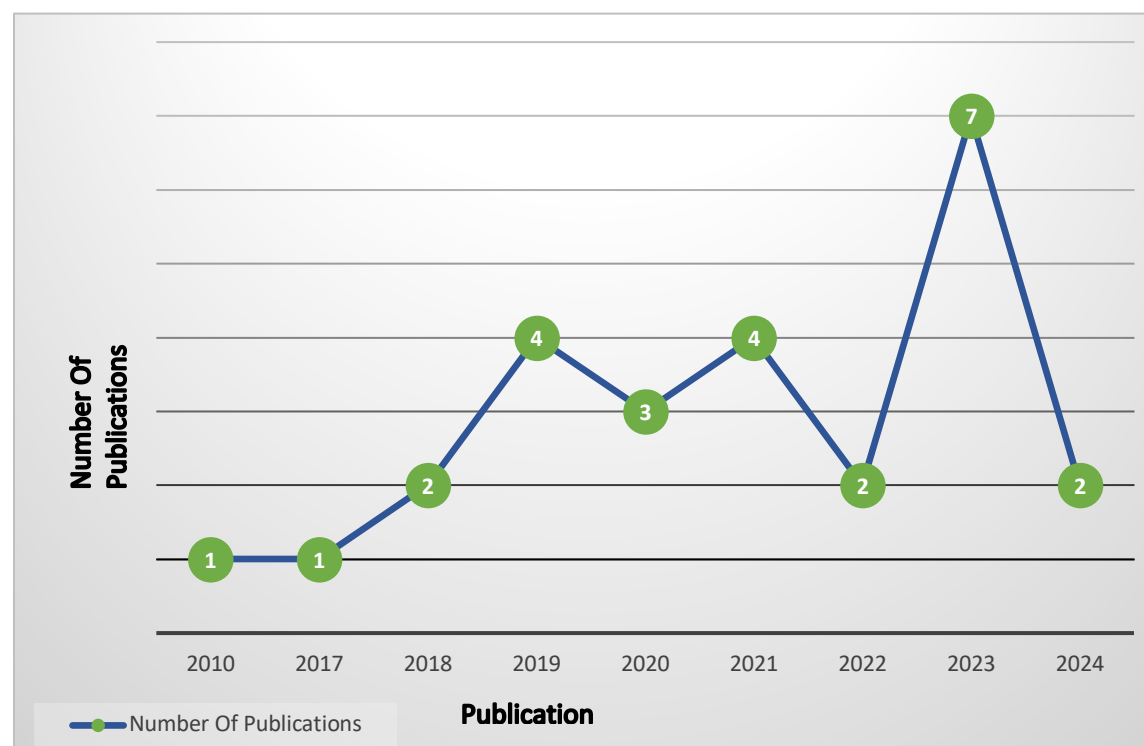
Seven studies have focused on employing *NLP for staging and quantification*. Lee SJ et al. <sup>16</sup> demonstrated the accurate interpretation of BT-RADS report scores through NLP trained on structured reports, while Zhang D et al. <sup>20</sup> utilized a rule-based algorithm to automatically categorize prostate MRI reports.

Six studies utilized NLP for *disease diagnosis*. Notably, in the studies by Dewaswala N et al. <sup>10</sup> and Liu Y et al. <sup>12</sup>, NLP effectively extracted diagnoses like hypertrophic cardiomyopathy and information about index lesions from breast MRI reports, respectively.

Moreover, NLP was integrated into four studies for *protocol selection*, exemplified by Trivedi H et al.<sup>25</sup> study, which employed natural language classification to identify contrast requirements in musculoskeletal MRIs.

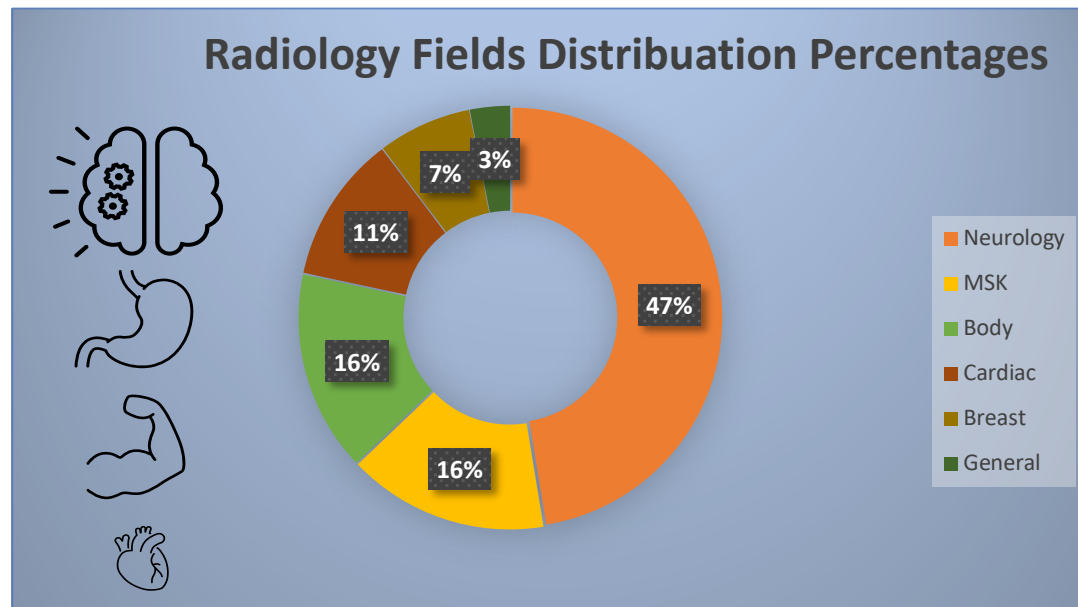
In the realm of *prognosis prediction*, Kim S et al.<sup>28</sup> introduced a novel computer model to analyze MRI reports of rectal cancer patients for estimating their survival times.

Additionally, Valtchiney VI et al.<sup>33</sup> explored methods for *identifying MRI safety risks* related to implanted devices in a study focusing on safety protocol compliance, while Chung EM et al.<sup>35</sup> not only developed a logistic regression model to predict patients requiring spine surgery but also investigated the use of ChatGPT technology to succinctly *summarize MRI reports* and enhance patient comprehension.

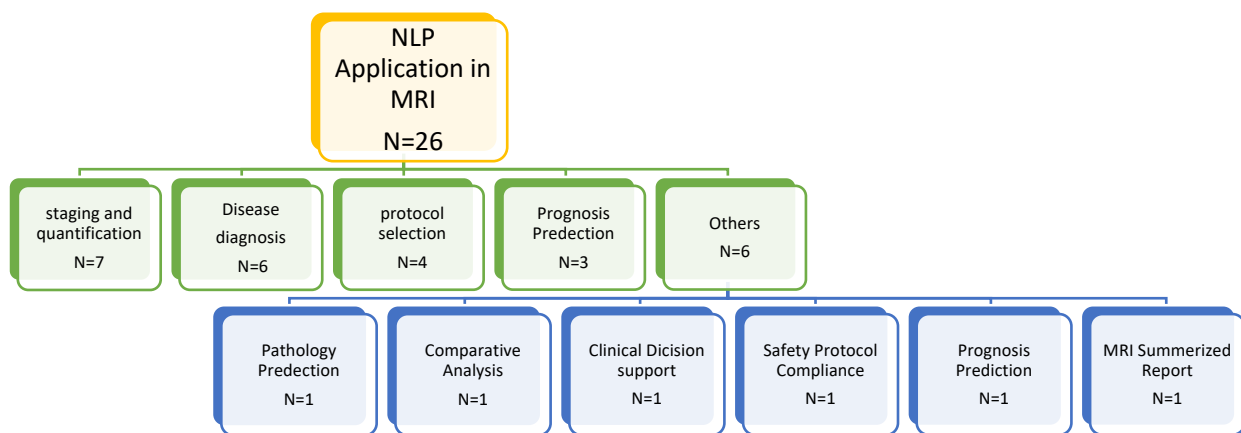


**Figure 2:** Evaluation of NLP applications in MRI research: trend analysis over time.

Note that the studies included were published before January 2024.



**Figure 3:** Radiology Fields Distribution Percentages



**Figure 4:** Natural language processing (NLP) MRI applications.

## ***Discussion***

Our review showcased the growing importance of NLP in MRI, with 26 publications covering over 160k MRI reports, across several organ systems. However, a key weakness is that only one prospective study was conducted <sup>27</sup>, such studies are essential for validating NLP algorithms in real-time clinical and operational settings to bolster clinical decision-making, workflow efficiency, and personalized medicine approaches. Future research should focus more on prospective studies to better validate the real-world benefits.

Additionally, recommendations include expanding NLP applications to other radiology fields such as pediatric radiology, and particularly in the context of cancer diagnosis and staging to broaden the impact and potential utility of NLP in the field.

The results of our review highlight several significant opportunities to streamline MRI imaging processes using advanced technology. By implementing more sophisticated protocols, bolstered by NLP, we could potentially reduce the time required for approving referrals <sup>26</sup>. This simplification means a more efficient workflow for healthcare providers.

Additionally, our findings suggest that NLP can simplify interpretative data for patients. This transparency allows patients to better understand their health information, which can improve their engagement and satisfaction with the treatment process <sup>35</sup>.

For radiologists, the technology we studied offers support in summarizing complex imaging results, such as those from MRI scans. This aid not only speeds up their workflow but also enhances the accuracy and comprehensiveness of the diagnostic data provided, especially in MRI reports with structured templates

<sup>19</sup>.

Looking ahead, the possibilities for further integration of this technology across different imaging domains are vast. Each specialty can learn from the others, leveraging technological advancements for the benefit of patient care and system efficiency.

Our review has limitations. First, we did not perform a meta-analysis due to the high heterogeneity in the methodologies and tasks used across the referenced studies, which made direct comparisons problematic. Second, our review was limited to articles sourced from PubMed, potentially omitting relevant studies published in other databases. Furthermore, the scope of our analysis was restricted to English-language publications, excluding potentially significant research available in other languages. Lastly, our findings are confined to the data available up to the point of our review, and as such, newer studies post-review are not considered.

In conclusion, NLP applications in MRI show potential for a change in the field. However, while the review revealed a wealth of evidence supporting the effectiveness of NLP in MRI analyses, the presence of just one prospective studies underscores the need for further research to validate NLP algorithms in real-time clinical and operational settings.

Moreover, expanding the scope of NLP usage to encompass other radiology specialties, presents opportunities for advancing healthcare practices. By addressing these recommendations in future research endeavors, the integration of NLP technologies stands to enhance clinical decision support, drive research advancements, and improve patient outcomes in the realm of MRI imaging.

## References

1. Sorin V, Klang E. Large language models and the emergence phenomena. *Eur J Radiol Open*. 2023;10:100494. doi:10.1016/j.ejro.2023.100494
2. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional Neural Networks for Radiologic Images: A Radiologist's Guide. *Radiology*. 2019;290(3):590-606. doi:10.1148/radiol.2018180547
3. Klang E. Deep learning and medical imaging. *J Thorac Dis*. 2018;10(3):1325-1328. doi:10.21037/jtd.2018.02.76
4. Sorin V, Barash Y, Konen E, Klang E. Deep Learning for Natural Language Processing in Radiology-Fundamentals and a Systematic Review. *J Am Coll Radiol JACR*. 2020;17(5):639-648. doi:10.1016/j.jacr.2019.12.026

5. Tessler I, Wolfowitz A, Livneh N, et al. Advancing Medical Practice with Artificial Intelligence: ChatGPT in Healthcare. *Isr Med Assoc J IMAJ*. 2024;26(2):80-85.
6. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. *Radiology*. 2016;279(2):329-343. doi:10.1148/radiol.16142770
7. Cai T, Giannopoulos AA, Yu S, et al. Natural Language Processing Technologies in Radiology Research and Clinical Applications. *Radiogr Rev Publ Radiol Soc N Am Inc*. 2016;36(1):176-191. doi:10.1148/rg.2016150080
8. Shah A, Aran S. A Review of Magnetic Resonance (MR) Safety: The Essentials to Patient Safety. *Cureus*. 2023;15(10):e47345. doi:10.7759/cureus.47345
9. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009
10. Dewaswala N, Chen D, Bhopalwala H, et al. Natural language processing for identification of hypertrophic cardiomyopathy patients from cardiac magnetic resonance reports. *BMC Med Inform Decis Mak*. 2022;22(1):272. doi:10.1186/s12911-022-02017-y
11. Kim C, Zhu V, Obeid J, Lenert L. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. *PloS One*. 2019;14(2):e0212778. doi:10.1371/journal.pone.0212778
12. Liu Y, Liu Q, Han C, Zhang X, Wang X. The implementation of natural language processing to extract index lesions from breast magnetic resonance imaging reports. *BMC Med Inform Decis Mak*. 2019;19(1):288. doi:10.1186/s12911-019-0997-3
13. Li MD, Deng F, Chang K, Kalpathy-Cramer J, Huang AJ. Automated Radiology-Arthroscopy Correlation of Knee Meniscal Tears Using Natural Language Processing Algorithms. *Acad Radiol*. 2022;29(4):479-487. doi:10.1016/j.acra.2021.01.017
14. Huhdanpaa HT, Tan WK, Rundell SD, et al. Using Natural Language Processing of Free-Text Radiology Reports to Identify Type 1 Modic Endplate Changes. *J Digit Imaging*. 2018;31(1):84-90. doi:10.1007/s10278-017-0013-3
15. Zaman S, Petri C, Vimalasvaran K, et al. Automatic Diagnosis Labeling of Cardiovascular MRI by Using Semisupervised Natural Language Processing of Text Reports. *Radiol Artif Intell*. 2022;4(1):e210085. doi:10.1148/ryai.210085
16. Lee SJ, Weinberg BD, Gore A, Banerjee I. A Scalable Natural Language Processing for Inferring BT-RADS Categorization from Unstructured Brain Magnetic Resonance Reports. *J Digit Imaging*. 2020;33(6):1393-1400. doi:10.1007/s10278-020-00350-0
17. Senders JT, Karhade AV, Cote DJ, et al. Natural Language Processing for Automated Quantification of Brain Metastases Reported in Free-Text Radiology Reports. *JCO Clin Cancer Inform*. 2019;3:1-9. doi:10.1200/CCI.18.00138



18. Cheng LTE, Zheng J, Savova GK, Erickson BJ. Discerning tumor status from unstructured MRI reports--completeness of information in existing reports and utility of automated natural language processing. *J Digit Imaging*. 2010;23(2):119-132. doi:10.1007/s10278-009-9215-7
19. Liu W, Cai L, Li Y. Application of natural language processing to post-structuring of rectal cancer MRI reports. *Clin Radiol*. 2024;79(2):e204-e210. doi:10.1016/j.crad.2023.10.032
20. Zhang D, Neely B, Lo JY, Patel BN, Hyslop T, Gupta RT. Utility of a Rule-Based Algorithm in the Assessment of Standardized Reporting in PI-RADS. *Acad Radiol*. 2023;30(6):1141-1147. doi:10.1016/j.acra.2022.06.024
21. Liu Y, Zhu LN, Liu Q, Han C, Zhang XD, Wang XY. Automatic extraction of imaging observation and assessment categories from breast magnetic resonance imaging reports with natural language processing. *Chin Med J (Engl)*. 2019;132(14):1673-1680. doi:10.1097/CM9.0000000000000301
22. Martín-Noguerol T, López-Úbeda P, Pons-Escoda A, Luna A. Natural language processing deep learning models for the differential between high-grade gliomas and metastasis: what if the key is how we report them? *Eur Radiol*. 2024;34(3):2113-2120. doi:10.1007/s00330-023-10202-4
23. Chillakuru YR, Munjal S, Laguna B, et al. Development and web deployment of an automated neuroradiology MRI protocoling tool with natural language processing. *BMC Med Inform Decis Mak*. 2021;21(1):213. doi:10.1186/s12911-021-01574-y
24. Brown AD, Marotta TR. A Natural Language Processing-based Model to Automate MRI Brain Protocol Selection and Prioritization. *Acad Radiol*. 2017;24(2):160-166. doi:10.1016/j.acra.2016.09.013
25. Trivedi H, Mesterhazy J, Laguna B, Vu T, Sohn JH. Automatic Determination of the Need for Intravenous Contrast in Musculoskeletal MRI Examinations Using IBM Watson's Natural Language Processing Algorithm. *J Digit Imaging*. 2018;31(2):245-251. doi:10.1007/s10278-017-0021-3
26. Eghbali N, Siegal D, Klochko C, Ghassemi MM. Automation of Protocoling Advanced MSK Examinations Using Natural Language Processing Techniques. *AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci*. 2023;2023:118-127.
27. Heo TS, Kim YS, Choi JM, et al. Prediction of Stroke Outcome Using Natural Language Processing-Based Machine Learning of Radiology Report of Brain MRI. *J Pers Med*. 2020;10(4):286. doi:10.3390/jpm10040286
28. Kim S, Lee CK, Choi Y, et al. Deep-Learning-Based Natural Language Processing of Serial Free-Text Radiological Reports for Predicting Rectal Cancer Patient Survival. *Front Oncol*. 2021;11:747250. doi:10.3389/fonc.2021.747250
29. Alskaf E, Frey SM, Scannell CM, et al. Machine learning outcome prediction using stress perfusion cardiac magnetic resonance reports and natural language processing of electronic health records. *Inform Med Unlocked*. 2024;44:101418. doi:10.1016/j.imu.2023.101418
30. Kim M, Ong KTI, Choi S, et al. Natural language processing to predict isocitrate dehydrogenase genotype in diffuse glioma using MR radiology reports. *Eur Radiol*. 2023;33(11):8017-8025. doi:10.1007/s00330-023-10061-z

31. Travis Caton M, Wiggins WF, Pomerantz SR, Andriole KP. Effects of age and sex on the distribution and symmetry of lumbar spinal and neural foraminal stenosis: a natural language processing analysis of 43,255 lumbar MRI reports. *Neuroradiology*. 2021;63(6):959-966. doi:10.1007/s00234-021-02670-6
32. Truhn D, Weber CD, Braun BJ, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. *Sci Rep*. 2023;13(1):20159. doi:10.1038/s41598-023-47500-2
33. Valtchinov VI, Lacson R, Wang A, Khorasani R. Comparing Artificial Intelligence Approaches to Retrieve Clinical Reports Documenting Implantable Devices Posing MRI Safety Risks. *J Am Coll Radiol JACR*. 2020;17(2):272-279. doi:10.1016/j.jacr.2019.07.018
34. Krebs B, Nataraj A, McCabe E, et al. Developing a triage predictive model for access to a spinal surgeon using clinical variables and natural language processing of radiology reports. *Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc*. Published online February 6, 2023. doi:10.1007/s00586-023-07552-4
35. Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. *Digit Health*. 2023;9:20552076231221620. doi:10.1177/20552076231221620