RWD23

Unlocking the Potential of Radiology Reports Using NLP: A Real-World Data Approach to Rotator Cuff Tear Severity

Or Shaked, MD MPH | Siegal Sadetzki, MD MPH | Talia Kustin, PhD | Ben Giladi, MD | Chen Patt, MD | Talia Tron, PhD | Briya, Tel Aviv, Israel

let's meet!

Objectives

Rotator cuff (RC) tears present significant clinical challenges, with treatment decisions—ranging from conservative management to surgical intervention—relying heavily on accurate **tear severity assessment**. However, International Classification of Diseases (ICD) codes often lack granularity to capture RC tear severity, complicating the use of **real-world data (RWD)** for research and clinical decision-making. This study aimed to utilize **machine learning (ML) Natural Language Processing (NLP)** to extract and classify RC tear severity from radiology reports, enabling scalable and precise identification of massive RC tears in RWD sources.

Methods

To address this, we developed an **on-premise** predictive model that categorizes **massive rotator cuff tears (MRCTs)** into three distinct types (**Table T1**) using hospital-collected shoulder MRI reports. Given computational constraints in clinical environments, we prioritize memory efficiency and explainability by employing a hybrid approach:

- **Rule-based extraction (regex)**
  Identifies key terms like "full-thickness", "supraspinatus", or "retraction" from reports
- **Classical NLP models (trainable on limited hardware)**
  Analyze structured features such as tear size, involved tendons, and atrophy descriptors
- **LLM integration (LLaMA 8B, deployed locally)**
  Generates contextual interpretations of ambiguous findings (e.g., "extensive tendinous involvement")

Table T1: **Definition of Massive Rotator Cuff Tear (MRCT):**
Any of the following conditions:

1	Specific mentioning of **massive** RC tear	**2**	Complete tear measuring ≥5 cm in any of the four RC tendons	**3**	Complete rupture of two or more of the four RC tendons

Shoulder MRI reports were collected from a single hospital system. Clinical experts annotated **623 MRI reports** for massive rotator cuff tear categories (**Table T1**). **147 reports (833 sentences)**, were further annotated at the **sentence level**, for tear type mentions (complete, massive, partial, no tear, cannot infer) and specific muscle involvement (supraspinatus, subscapularis, infraspinatus, teres minor).
These were used for the **Briya© NLP model training**.

We developed a **hybrid pipeline**. MRI reports were first scanned for **relevant keywords**, then analyzed by **specialized classifiers**. The **AI model** provided additional insights, and results were merged using predefined rules to determine if a tear was "massive" (***Table T2*** *and* ***Figure A***). This approach balances **accuracy** with the ability to run **securely** within hospital systems while maintaining **transparent decision-making**. Performance was assessed separately for **individual NLP models** and the **integrated pipeline** using standard classification metrics (**accuracy, precision, f1score**).
Clinical experts reviewed outputs to validate clinical relevance and explainability of predictions.

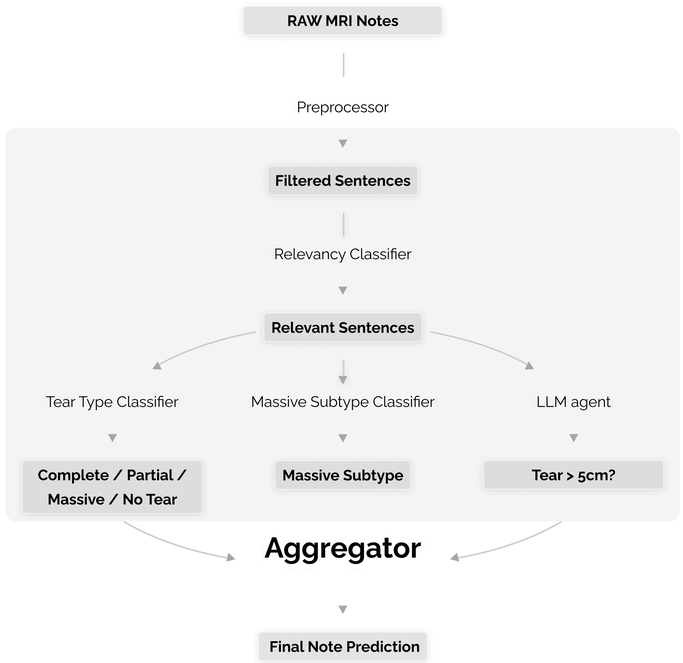
Table T2: **Briya NLP Model Pipeline**

Pipeline Stage	Description	Technical Details
Text Preprocessing	Segments MRI reports into sentences and filters content using shoulder pathology keywords (see below)	Hebrew token normalization, split to sentences keyword filtering*
1. Relevancy classifier	Uses ML to classify sentences as clinically relevant/irrelevant for rotator cuff tear descriptors	Binary classifier**
2. Tear Type Categorization	Classifies eligible sentences into discrete rotator cuff tear categories	Multiclass classifier** to **one of the** following: complete/ massive/partial/no tear/indeterminate
3. Massive Tear Subtyping	Refines massive tear predictions using LLMs and NLP for contextual analysis	Multilabel classifier** to **any of the following** (complete, massive 1, massive 2)
Result Aggregation	Synthesizes predictions across all stages into unified report-level classification	Aggregation logic combining pipeline outputs

* **Keywords filter for task**: Shoulder, muscle, rotator, retraction, thickness, tendon, tear, raptor
** Based on logistic regression with l2 regularization & tf-idf vectorizer, ngrams=(2,4)

Briya NLP Model Pipeline

Figure A:



LLM* to identify tears larger than 5cm:

```
massive_tear2_template = """Does the following text mention a shoulder tear larger then 5 cm?
INSTRUCTIONS:
- If a shoulder tear size is mentioned (ארוך, גדול, קרוח) - return "larger than 5cm" or "smaller than 5cm"
- If no shoulder tear is mentioned -  return "not mentioned"
- If a shoulder tear is mentioned but with no explicit size  - "not mentioned"

Response should be in the following format:
explanation = short explanation of the conclusion
result = "larger than 5cm" \ "smaller than 5cm" \ not mentioned

TEXT:
$NOTE
}
"""
```

Hebrew Normalization:

```
def normalize_hebrew(text):
    # 1. Remove Hebrew diacritics (Nikud and Taamim)
    text = re.sub(r'[\u0591-\u05C7]', '', text)

    # 2. Remove non-informative prefixes (adjust as needed)
    text = re.sub(r'\b[וכבלמה]{1}(?=[\u05D0-\u05EA])', '', text)

    # 3. Remove non-informative suffixes (adjust as needed)
    text = re.sub(r'([\u05D0-\u05EA]+)(ים|ות|יה|ית|ה)\b', r'\1', text)

    # 4. Remove unwanted characters (keep numbers, Hebrew, English)
    text = re.sub(r'[^a-zA-Z0-9\u05D0-\u05EA\n\r,. ]+', '', text)

    # 5. Normalize whitespace
    text = re.sub(r'\s+', ' ', text).strip()

    return text
```

Results

Radiology reports of **406 patients** were analyzed,
with a **mean age of 58.8 years** (standard deviation±8.78),
and 61.6% of the patients were male.
56.4% of the scans were of the right shoulder and 43.6% of the left shoulder

Data preprocessing:

Report Level:
A total of 623 reports were annotated, with 619 reports (99%) containing relevant keywords. Based on clinical expert tagging, **97 cases (15.5%) were identified as massive rotator cuff tears**.

Sentence Level:
Reports consisted of 24,690 sentences in total, with 6,392 (25%) relevant sentences after filtering related keywords. The mean sentence per report was 20 with standard deviation of ±5 sentences.

Data Descriptions:

We used a dual annotation strategy: Combines **document-level classification** with **fine-grained sequence labeling (by sentence):**

Training: Sentence-level annotations
- 147 reports with sentence-level annotations (833 sentences) used for model training
- Used for training sequence labeling models

Evaluation: Full-report annotations
- 623 reports fully annotated for report-level classification (e.g., 15% massive tears)
- 99% of reports contain relevant keywords
- Average report length: 20 ± 5 sentences

Sentence-level annotations
Label distribution by sentence:

Label	Count	Proportion
For Sure Complete*	321	0.38
Not Related	263	0.31
Partial	143	0.17
No Tear	47	0.05
For Sure Massive_1	25	0.03
For Sure Massive_2	9	0.01
Other	37	0.04

*In one of the four RC tendons - Supraspinatus, Infraspinatus, Teres minor, Subscapularis

Briya NLP model:
Results

Pipeline Stage	Description
Text Preprocessing	• Notes from 623 MRI reports - 527 not massive, 97 massive • Split to train (60%) and test (40%) for training and evaluation • Split to 24,690 sentences • 25% (6,392 sentences) filtered by relevant keywords
Relevancy classifier	• Average Precision: 91% • Average recall: 96%
Final note prediction	• Average Precision: 91% • Average Recall: 90%

Conclusions

This RWD-driven NLP approach offers a robust, scalable solution for identifying and classifying massive RC tears from MRI radiology reports, addressing the limitations of severity assessment of ICD coded conditions in real-world healthcare settings.

By enabling enhanced data extraction and improving the precision of RC tear diagnosis, this method has the potential to support both clinical decision-making and large-scale RWD research on RC tears.