NLP SUMMIT HEALTHCARE

Extracting Disease Severity Measures from Unstructured Clinical Notes Using Spark NLP

Vikas Kumar

Senior Data Scientist
OMNY Health
vikas@omnyhealth.com

Lawrence Rasouliyan

Head, Biostatistics & Data Science
OMNY Health
lawrence@omnyhealth.com

OMNY HEALTH

Talk Outline

I. Disease Severity Measures: Background and Clinical Importance (Lawrence)

II. Disease Severity Measures and Spark NLP (Vikas)

Disease Severity Measures: Background and Clinical Importance

Disease Severity Measures: Overview

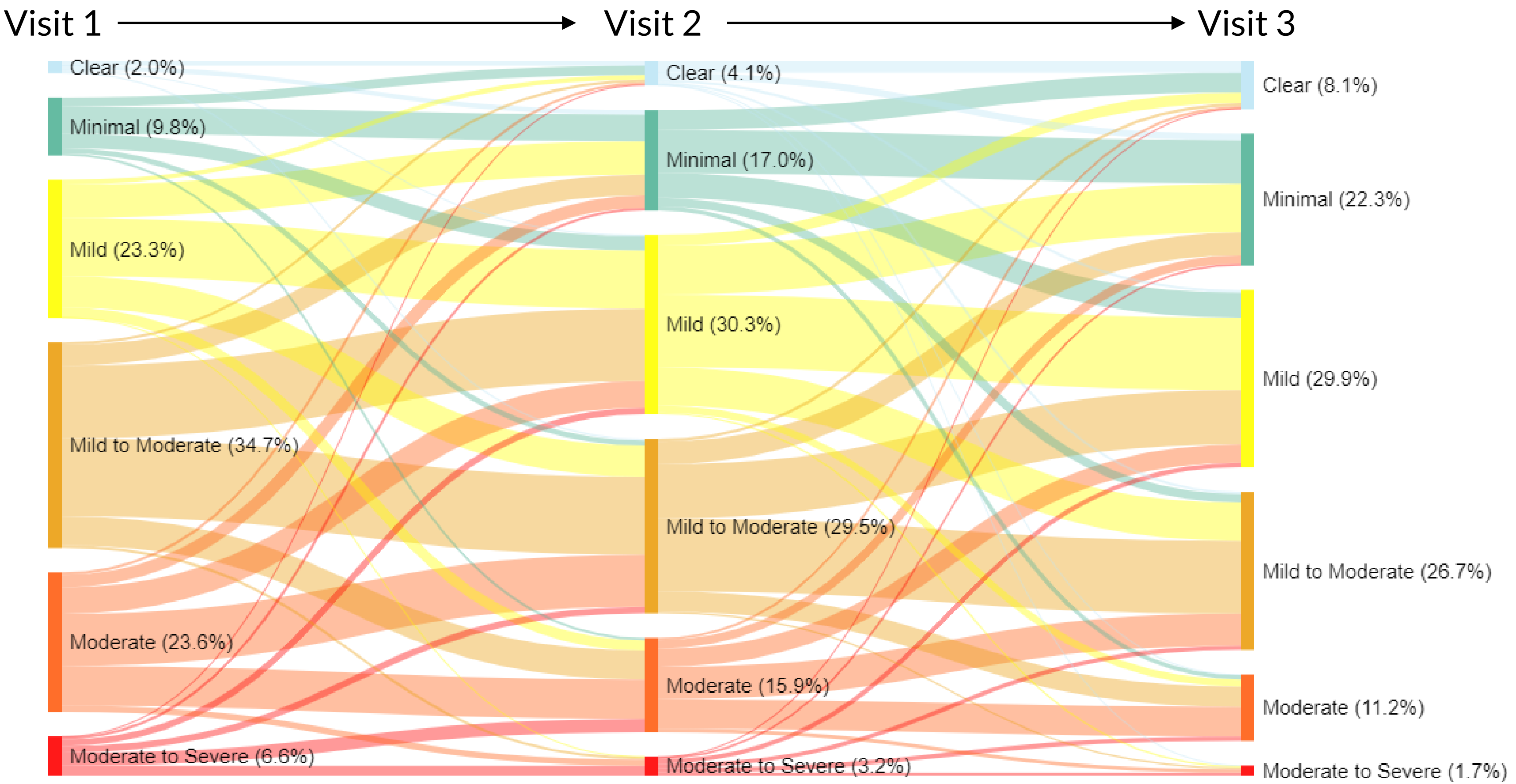
- Disease severity is a broad term to describe the current state of a particular medical condition.
- Usually related to how the patient feels and/or symptoms related to the disease
- Severity determines the management of almost all diseases.

Structured Data Example from Acne Vulgaris

In your experience, among all patients you have seen with this condition, how severe is this patient's condition?

- □ Clear
- □ Minimal
- □ Mild
- □ Mild to Moderate
- □ Moderate
- □ Moderate to Severe
- □ Severe

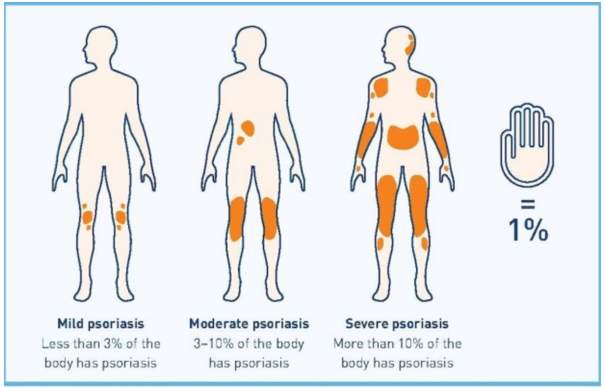
Structured Data Example from Acne Vulgaris



Disease Severity Measures: Overview

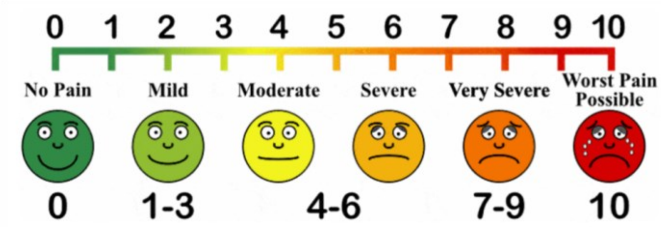
- Disease severity can be measured by many factors such as
  - Symptoms
  - Physical appearance
  - Extent
  - Complications
  - Laboratory/biometric values
  - Other factors or combinations
- Disease severity measures can be
  - Objective or subjective
  - Physician assessed or patient reported
  - Absolute or relative

Disease Severity Measures: Examples

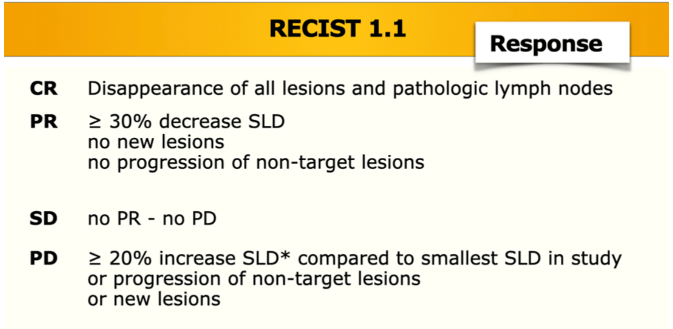

= 1%
Mild psoriasis
Less than 3% of the body has psoriasis
Moderate psoriasis
3–10% of the body has psoriasis
Severe psoriasis
More than 10% of the body has psoriasis

©2020 National Psoriasis Foundation

https://www.psoriasis.org/why-treat/



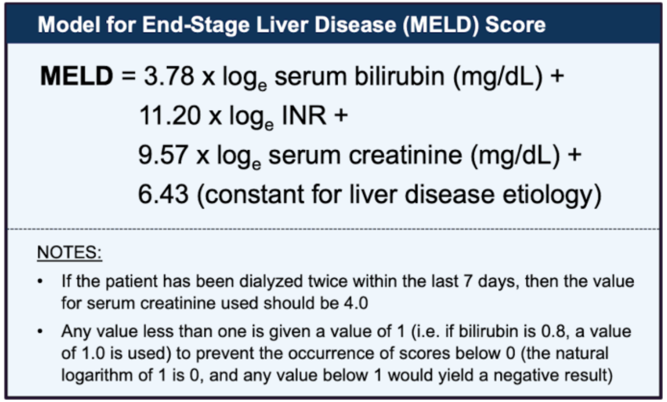
https://greatbrook.com/visual-analog-survey-scale/



RECIST 1.1 **Response**

CR	Disappearance of all lesions and pathologic lymph nodes
PR	≥ 30% decrease SLD no new lesions no progression of non-target lesions
SD	no PR - no PD
PD	≥ 20% increase SLD* compared to smallest SLD in study or progression of non-target lesions or new lesions

https://radiologyassistant.nl/more/recist-1-1/recist-1-1-1



Model for End-Stage Liver Disease (MELD) Score

$$\textbf{MELD} = 3.78 \times \log_e \text{serum bilirubin (mg/dL)} + 11.20 \times \log_e \text{INR} + 9.57 \times \log_e \text{serum creatinine (mg/dL)} + 6.43 \text{ (constant for liver disease etiology)}$$

NOTES:

- If the patient has been dialyzed twice within the last 7 days, then the value for serum creatinine used should be 4.0
- Any value less than one is given a value of 1 (i.e. if bilirubin is 0.8, a value of 1.0 is used) to prevent the occurrence of scores below 0 (the natural logarithm of 1 is 0, and any value below 1 would yield a negative result)

https://www.hepatitisc.uw.edu/go/evaluation-staging-monitoring/evaluation-prognosis-cirrhosis/core-concept/all

Disease Severity Measures: Clinical Research Importance

- Patient improvement/worsening
- Risk stratification
- Treatment strategies
- Effectiveness of treatment
- Related to quality of life
- Associated with healthcare resource utilization and costs

Disease Severity Measures: Setting Up the Problem

- Disease severity measures are often missing in structured EHR data
- The extent to which they are documented in the unstructured clinical notes may vary by practice, provider, disease, and severity measure
- A reliable method of applying NLP to the unstructured clinical notes to extract disease severity measures would be beneficial for disease management and retrospective clinical research.

Disease Severity Measures and Spark NLP

Objective

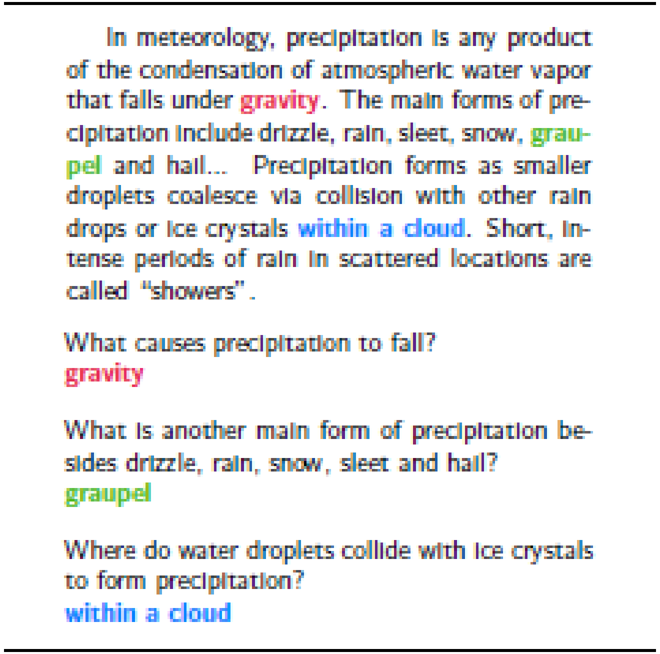
- To use a pretrained, transformer-based QA model to extract severity scores from unstructured clinical notes.

Background: Question-answering and SQuAD

- Question-answering (QA) is a NLP subtask in which a model answers questions about a context.

- In closed-domain QA, the answer must appear in the context for the question to be answered.

- There is an abundance of language models trained on QA datasets that are open-source.



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

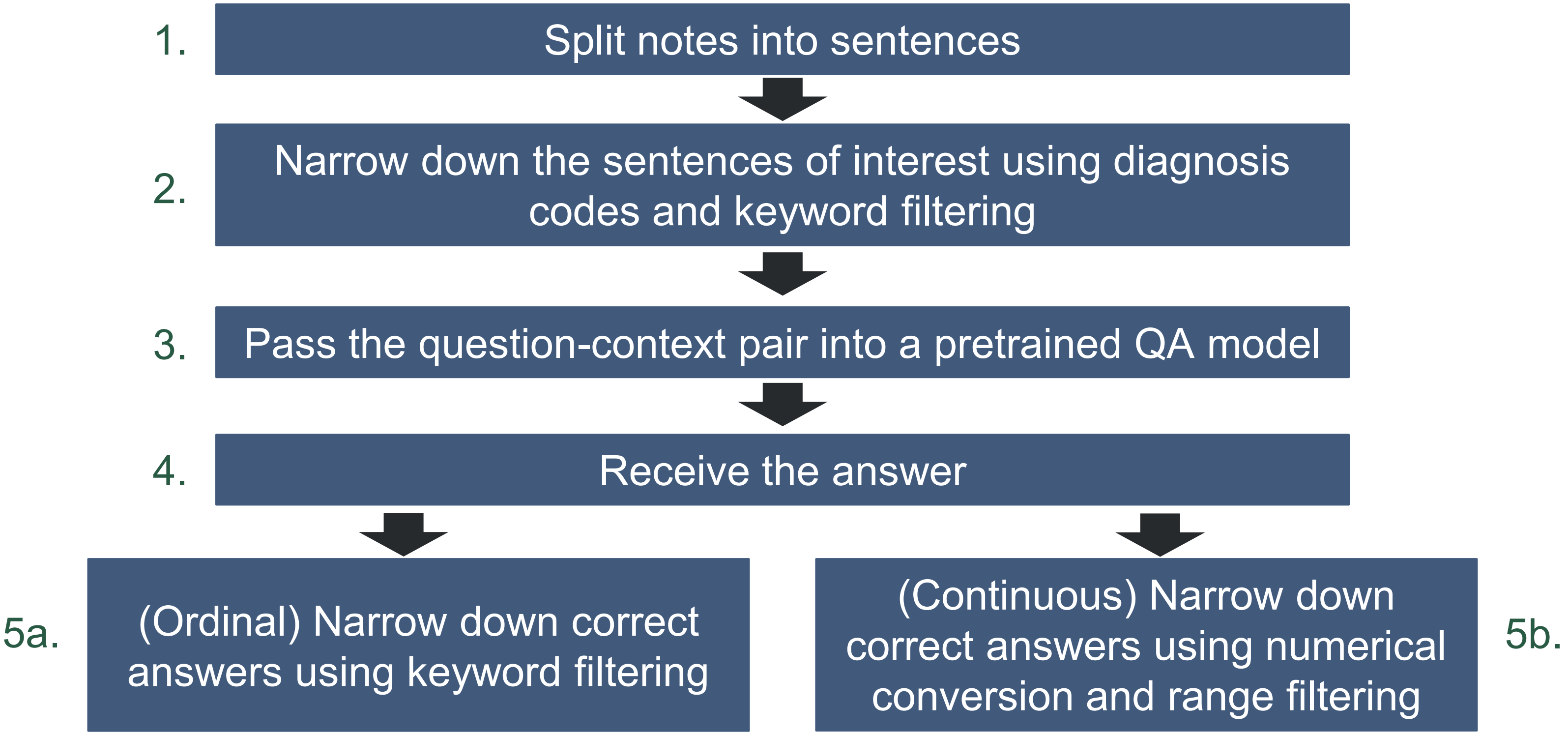
Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

Rajpurkar et al., 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. https://arxiv.org/pdf/1606.05250.pdf.

Formulating Question-Context Pairs

Measure	Diagnosis Code	Question	(Search strings, ranges, answer strings, etc.)
Ejection fraction	I50	What is the EF?	...
Abdominal pain	R10.9	What is the severity of the abdominal pain?	...
Lower back pain	M54.5	What is the severity of the back pain?	...
...	...	...	...

Solution Components



1. Split notes into sentences

2. Narrow down the sentences of interest using diagnosis codes and keyword filtering

3. Pass the question-context pair into a pretrained QA model

4. Receive the answer

5a. (Ordinal) Narrow down correct answers using keyword filtering

5b. (Continuous) Narrow down correct answers using numerical conversion and range filtering

A continuous example

- Context:

```
Co morbidities pulmonary edema, esrd, htn, copd, dm, chf ef
40%,
```

✓ Patient has a diagnosis code of I50.9

✓ Sentence contains one of the search strings (“ef”)

- Question: “What is the EF?”
- Result: “40%” (string)
- Final result after casting and checking range: 40.0 (float)

An ordinal example

- Context

```
<AGE> F here w/ mom for nausea/vomting.  Had some mild
generaliazed intermittent ab pain over past several days  This
evening she had vomiting w/ dinner.  exudate or posterior
oropharyngeal erythema.  Additionally patient is eating
crackers in the room without any difficulty.
```

✓ Patient has a diagnosis code of R10.9

✓ Sentence contains one of the search strings (“mild”)

- Question: “What is the severity of the abdominal pain?”
- Result: “mild generaliazed intermittent” (string)
- Final result: mild (categorical string) (since answer contains “mild”)

Spark NLP Pipeline Components

```
MODEL_NAME = 'twmkn9/albert-base-v2-squad2’



document_assembler = MultiDocumentAssembler() \
...

spanClassifier = AlbertForQuestionAnswering.loadSavedModel(
     '{}/saved_model/'.format(MODEL_NAME),
     spark
) \
...


pipeline = Pipeline().setStages([
    document_assembler,
    spanClassifier
])
```

Defining the model

DocumentAssembler()

SpanClassifier

- Load a pretrained HuggingFace model into Spark

Initializing the Pipeline Object

Results: Severity Scores Detected

```
+-------------------------------------------+-----+
|QS_NAME                                    |count|
+-------------------------------------------+-----+
|Congestive heart failure: Ejection fraction|10000|
|Abdominal Pain: Severity                   |7934 |
|Low Back Pain: Severity                    |3952 |
|CKD Stage                                  |2949 |
|Psoriasis: BSA                             |1598 |
|Depression: PHQ-9                          |1264 |
|...(17 other measures implemented)...      |...  |
+-------------------------------------------+-----+
```

Results: Accuracy for Selected Measures

```
+-----------------------+------------+--------------+-----------------+------+
|QS_NAME                |sample_total|sample_correct|overall_total_est|acc   |
+-----------------------+------------+--------------+-----------------+------+
|Psoriasis: PASI        |1           |1             |20               |100.0 |
|Derm: Fitzpatrick scale|1           |1             |20               |100.0 |
|Psoriasis: BSA         |74          |66            |1480             |89.2  |
|Hidradenitis supp: Hurl|36          |32            |720              |88.9  |
|Atopic dermatitis: BSA |21          |14            |420              |66.7  |
+-----------------------+------------+--------------+-----------------+------+
```

Limitations

- Accuracy could be improved
  - Accuracy highly dependent on severity measure
  - Multiple severity scores in the same sentence may lead to errors
- Recall could be improved
  - Not all keyword synonyms are accounted for
- Ranges of severity scores not optimally handled

Future Directions

- Domain adaptation and fine-tuning of model
- Implementation of additional severity score measures across various treatment areas

OMNY HEALTH

Thank you!

Vikas Kumar

Senior Data Scientist

OMNY Health

vikas@omnyhealth.com

Lawrence Rasouliyan

Head, Biostatistics & Data Science

OMNY Health

lawrence@omnyhealth.com

NLP SUMMIT HEALTHCARE

Presented by

John Snow LABS

www.nlpsummit.org