Extracting concepts from Dutch clinical text

Concept extraction from Dutch clinical text

PRESENTER: **Tom Seinen**

INTRO:

- EHR databases contain **vast amounts** of unstructured **text data.**
- Free-text cannot be directly used for analysis.
- **Named-Enity-Recognition (NER)** is the task of extracting clinical concept from the free-text.
- **No open-source NER tools** exist for concept extraction from **Dutch clinical text**.
- We created and evaluated an open-source extraction tool for the extraction of concepts from **Dutch clinical text** by converting an existing framework, **MedSpacy**.

METHODS

Dataset – Dutch GP database with 2.8 million patients (IPCI) from 1992 to 2022, converted to the OMOP CDM.

Text preprocessing – Only keep alphanumeric characters, tokeninize with *Dutch SpaCy model.*

Concept extraction – MedSpacy's quickUMLS using the *Dutch SNOMED CT ontology*.

Context extraction – MedSpacy's context extraction using *Dutch target rules.*

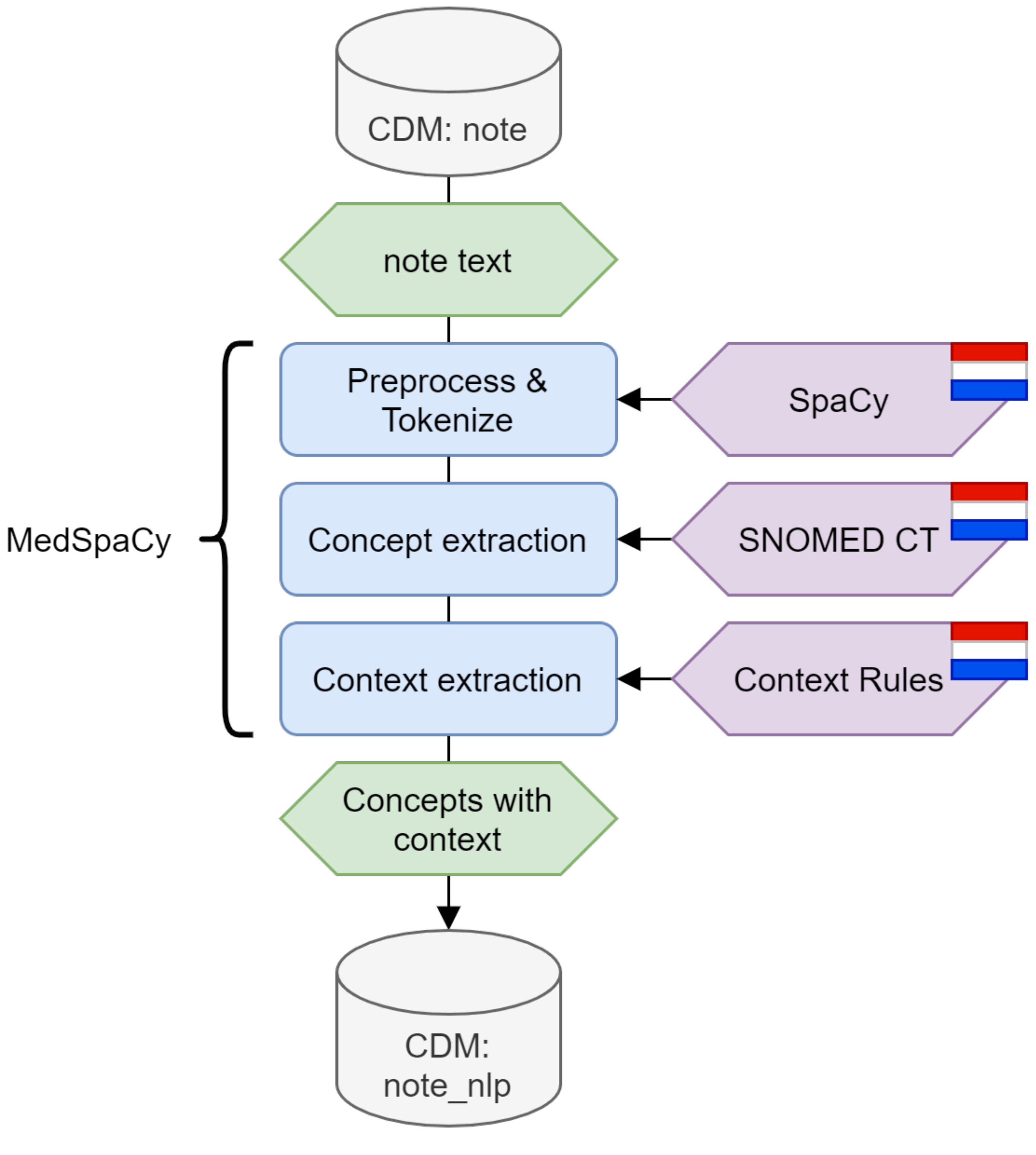
Exploratory setup – Framework was applied to notes surrounding the occurrences of 6 specific coded conditions. A window was of 1 day before and 1 day after the code occurrence. The most important concepts were identified for each code using the TFIDF value.

RESULTS

	Alzheimer's disease	Depressive disorder	Infection disease of cardiovascular system
# code occurrences	1.226.047	5.581.176	122.482
# codes per patient	40,4	31,2	18,3
Mean # notes per code occurrence	4,7	4,0	4,8
Median; mean # words per note	7; 19,0	6; 15,0	4; 15,5
# concepts	4.216.892	14.702.150	452.449
# unique concepts	10.590	15.713	5.991
% negated concepts	16,1%	13,7%	14.8%
% historical concepts	3,5%	5,0%	5,5%
Mean # extracted concepts per code occurrence	20,2	13,6	21,9
Mean # extracted concepts per note	7,5	5,6	7,9
Ratio extracted concepts / note size	0,39	0,37	0,51

Summarizing statistics over 3 condition codes

Text processing pipeline



CDM: note

note text

MedSpaCy

Preprocess & Tokenize

SpaCy

Concept extraction

SNOMED CT

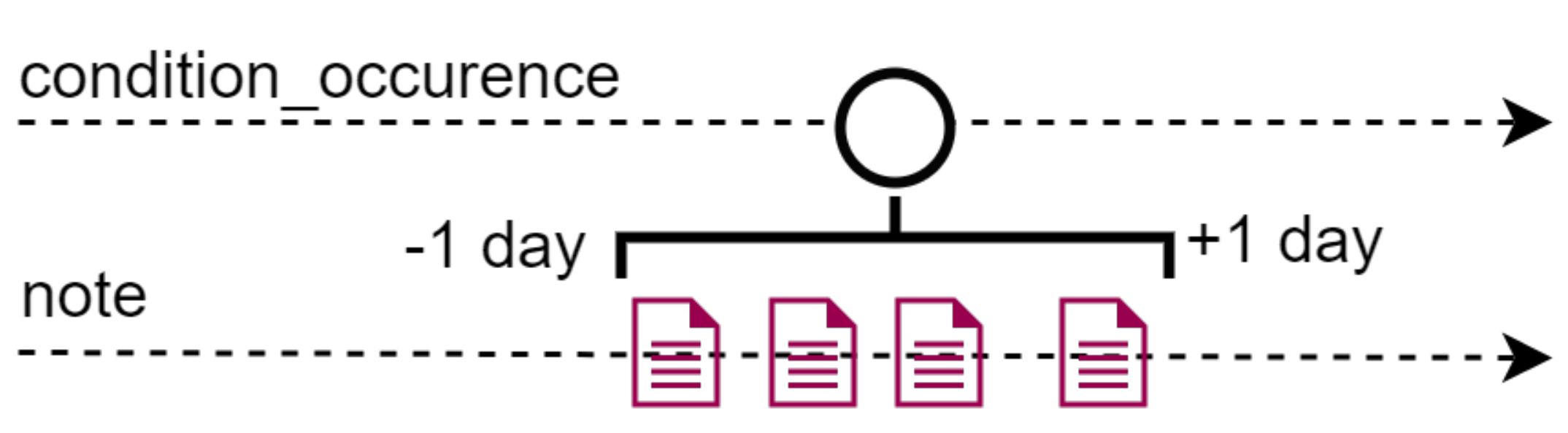
Context extraction

Context Rules

Concepts with context

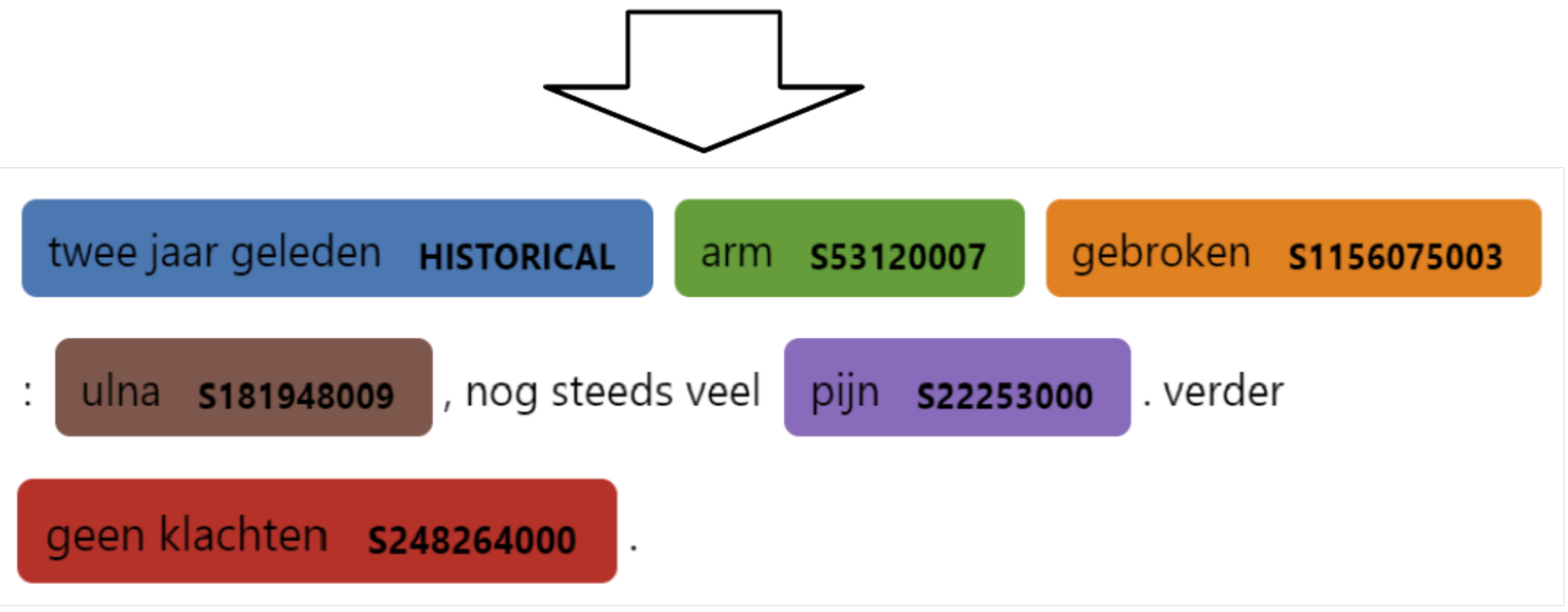
CDM: note_nlp

Experimental setup

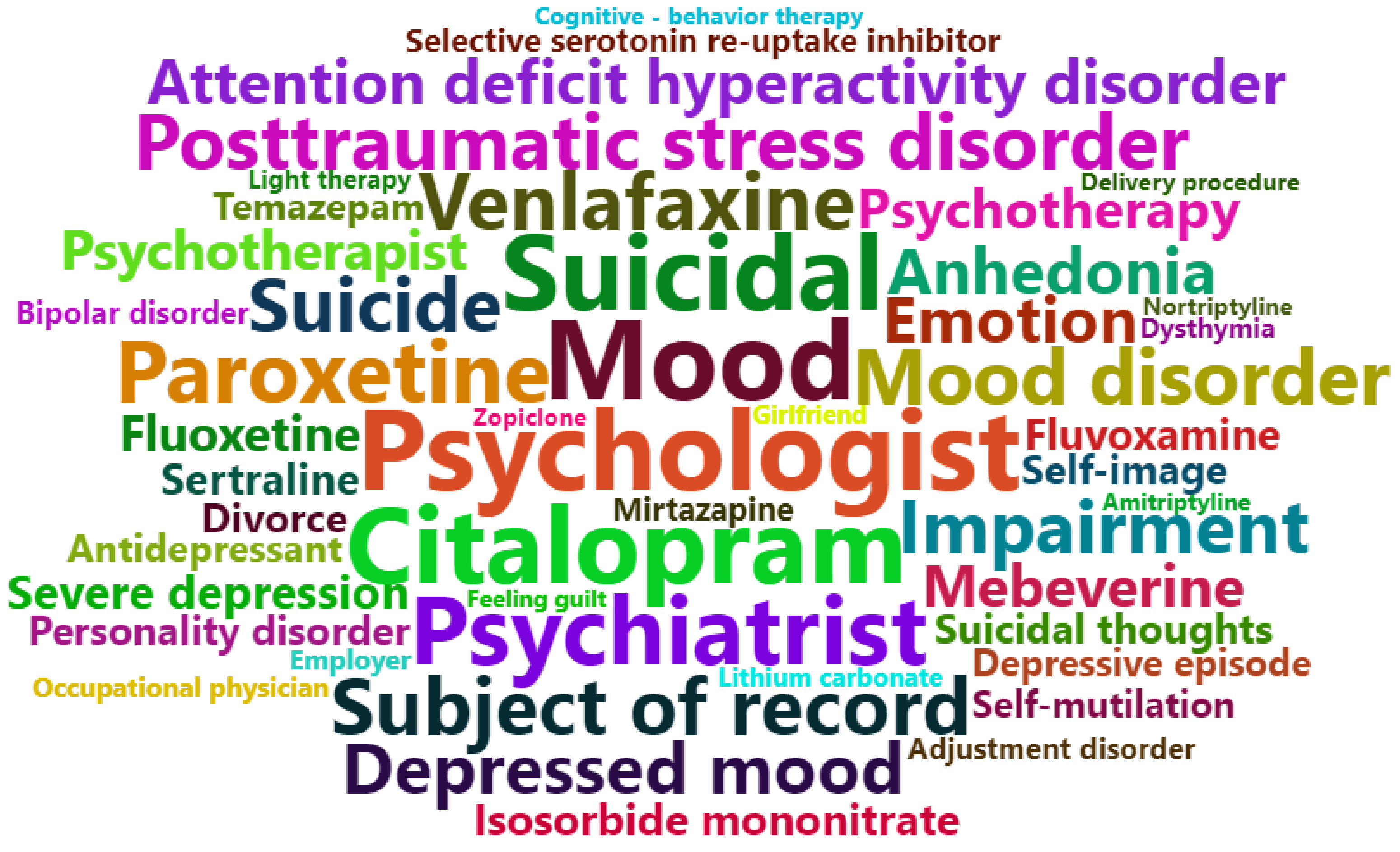


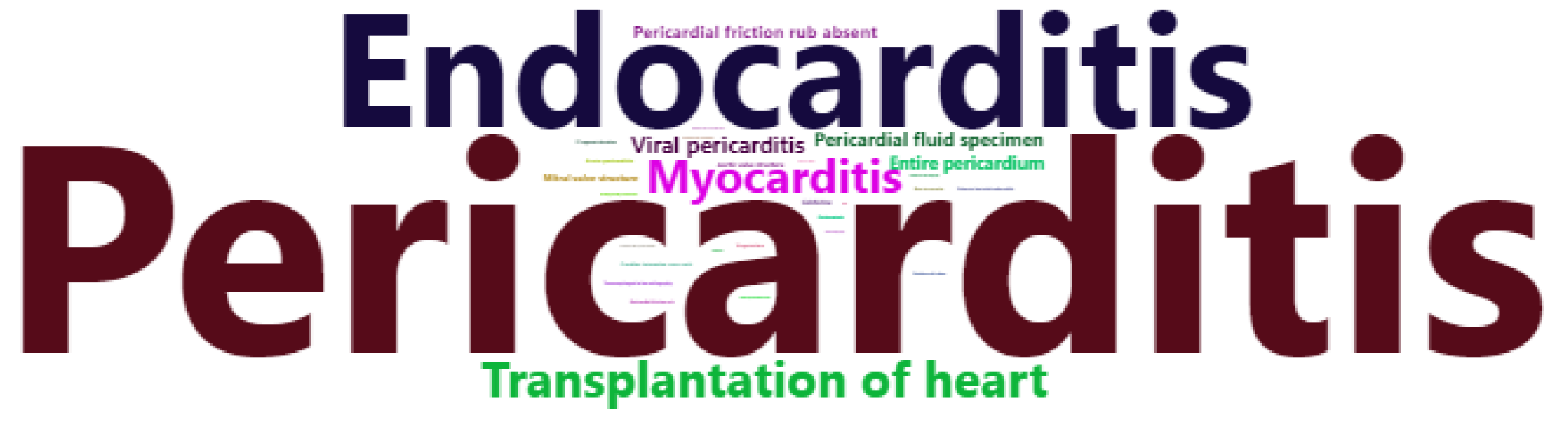
Concept extraction example

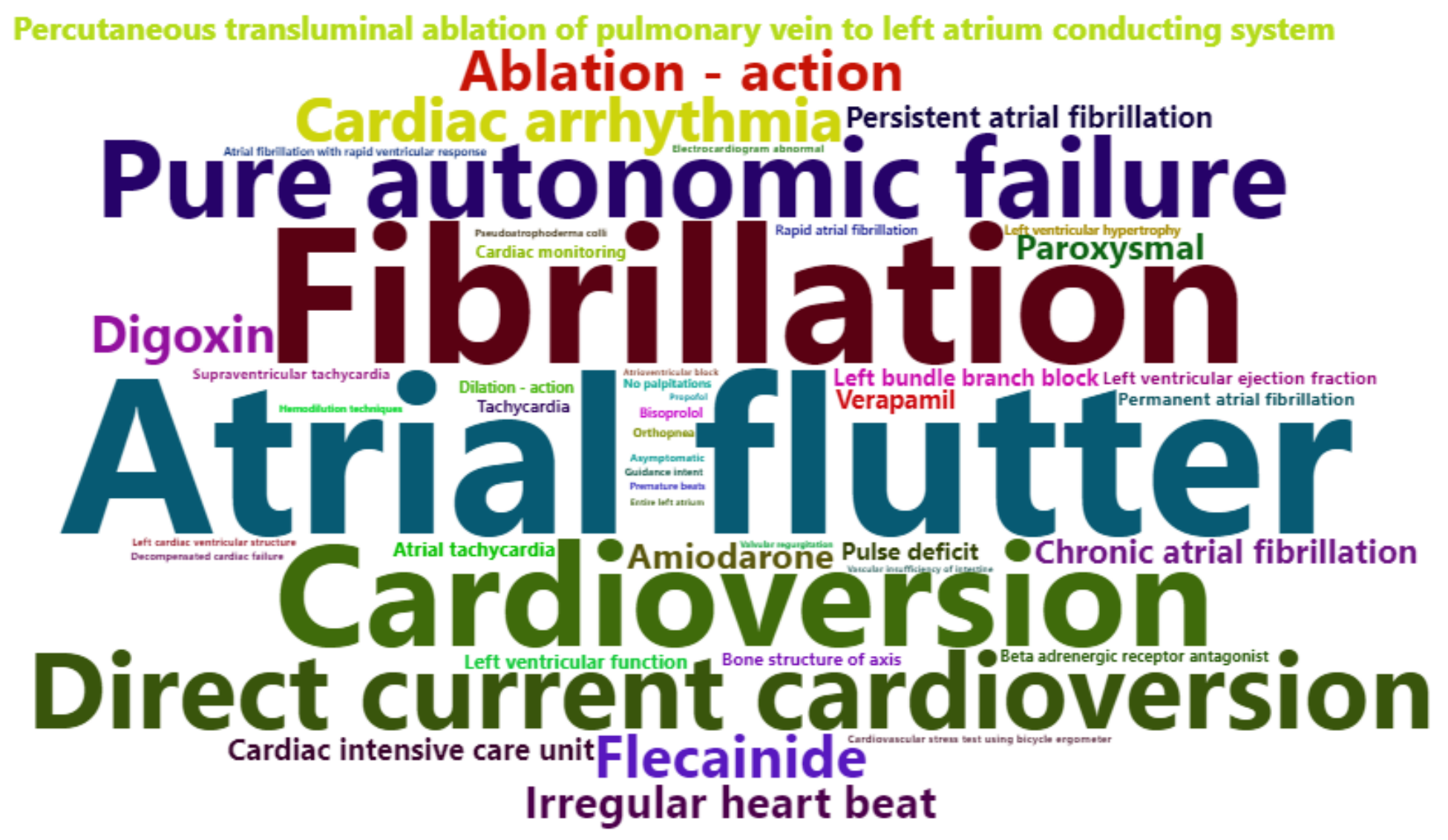
```
twee jaar geleden arm gebroken:
ulna, nog steeds veel pijn. verder
geen klachten.
```



twee jaar geleden **HISTORICAL** arm **S53120007** gebroken **S1156075003** : ulna **S181948009** , nog steeds veel pijn **S22253000** . verder geen klachten **S248264000** .







Top 5 most important concepts (by TFIDF) per condition

	Atrial Fibrillation /flutter	Asthma	Alzheimer's disease
1	Atrial flutter	Spirometry	Dementia
2	Fibrillation	Forced expired volume	Clonidine
3	Cardioversion	Forced vital capacity	Alzheimer's disease
4	Paroxysmal atrial fibrillation	Pulmonary function	Case manager
5	Pure autonomic failure	Exercise-induced asthma	Type 2 diabetes mellitus

	Depressive disorder	Infection disease of cardiovascular system	Fracture ulna/radius
1	Mental health care	Pericarditis	Bone structure of distal radius
2	Psychologist	Endocarditis	Entire radius
3	Mood	Transplantation of heart	Entire ulna
4	Citalopram	Myocarditis	Fracture
5	Suicidal	Viral pericarditis	Reduction - action

CONCLUSION

- We **analyzed concepts** around 6 coded condition occurrences in a Dutch OMOP CDM.
- The found concepts are **descriptive and informative** of the coded conditions.
- The extracted concepts show the **ambiguity** of several **ICPC codes.**
- The **detailed information** extracted from the free-text can be used **in further research or to improve the ETL to OMOP**

FUTURE STEPS

- **Quantitative evaluation** and validation of the concept extraction framework
- **Use of data** in :
  - **Patient level prediction**
  - **Diagnostic classification**
  - **ETL to OMOP CDM**
- Effects of **spelling correction** on concept extraction
- **Compare extracted concepts** with the **structured data** in the OMOP CDM.

Tom M. Seinen[1], Jan A. Kors[1], Erik M. van Mulligen[1], Peter R. Rijnbeek[1], [1]Department of Medical Informatics, Erasmus MC, Rotterdam, The Netherlands

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.