INTRO:

- EHR databases contain vast amounts of unstructured text data.
- Free-text cannot be directly used for analysis.
- Named Entity 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, tokenized 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

- Top 5 most important concepts (by TFIDF) per condition:
  - **Atrial Fibrillation**/Flutter
  - Asthma
  - Alzheimer's disease
  - Depression
  - Chronic obstructive pulmonary disease

- Summary statistics over 3 condition codes:
  - **Alzheimer's disease**
    - Code occurrences: 2,296,74
    - Mean # notes per code occur: 8.7
    - Mean # concepts per note: 5.2
    - % negated concepts: 12.3%
    - % historical concepts: 3.5%
  - **Depression**
    - Code occurrences: 1,342,846
    - Mean # notes per code occur: 5.8
    - Mean # concepts per note: 4.3
    - % negated concepts: 13.7%
    - % historical concepts: 4.3%
  - **Chronic obstructive pulmonary disease**
    - Code occurrences: 1,040,653
    - Mean # notes per code occur: 4.4
    - Mean # concepts per note: 3.8
    - % negated concepts: 14.7%
    - % historical concepts: 4.3%

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. Seinen1, Jan A. Kors1, Erik M. van Mulligen1, Peter R. Rijnbeek1, 1Department of Medical Informatics, Erasmus MC, Rotterdam, The Netherlands