Use of unstructured text data in electronic health records to improve patient-level prediction models

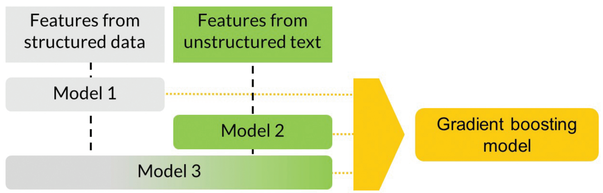
PRESENTER: **Tom M. Seinen**

INTRO:

- Lots of unstructured text available in OMOP cdm databases.
- Clinical text possibly contains additional/other information compared to structured/coded data.
- Use this information in PLP models.
- **Contributions**:
  - Customizable language independent NLP pipeline for within the OHDSI framework
  - Example study on a Dutch OMOP cdm database
- **Objective**:
  - Explore the contribution of features extracted from clinical text to the development of patient-level prediction models.

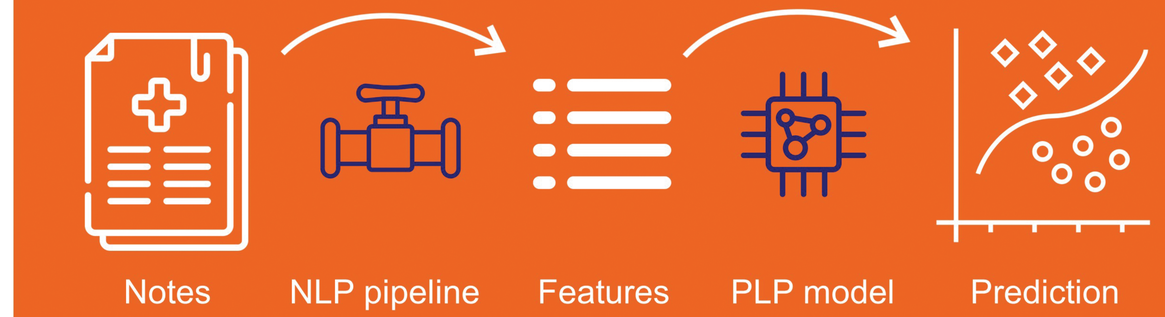
METHODS

- **Natural language processing pipeline**
  1. Retrieve cohort notes from CDM
  2. Pre-process note text
  3. Tokenize note text
  4. Vectorize text: Numeric representation
- **Proof of concept study**
  - Database: ***Integrated Primary Care Information (IPCI)***
  - Target: ***Type 2 diabetes adult patients (16,437)***
  - Outcome: ***30-day risk of heart failure (92)***
- **Features/Covariates:**
  - Observation time: ***1 year***
  - Structured data:
    - ***All FeatureExtraction covariates*** (30d/365d)
  - Unstructured text:
    - ***1 and/or 2 ngrams, bag-of-words(TFIDF)*** (365d)
- **Experimental setup**



RESULTS

Model	AUC (CI)	AUPRC
1 Structured	0.67 (0.56-0.79)	0.012
2a Text n=1	0.68 (0.56-0.97)	0.011
2b Text n=1,2	0.69 (0.58-0.80)	0.012
3a Struct. + Text n=1	0.80 (0.73-0.88)	0.022
3b Struct. + Text n=1,2	0.71 (0.59-0.82)	0.017



Feature extraction from unstructured **clinical text** for **patient-level prediction**

TRITON: Text Represented In Terms Of Numeric-features
https://github.com/mi-erasmusmc/Triton

Model settings

Split:	75% training, 25% test.
Gradient boosting:	maxDepth: 17; minRows=2; ntrees:100; learnRate:0.1
Hyperparameter tuning:	3-fold cross-validation

NLP pipeline settings:

Preprocessing:	Lowercase; digit and symbol removal
Tokenization:	Word tokenization (stringi)
Stopword removal:	Dutch stopwords (SnowballC)
Word ngrams	Uni and bigram (n=1,2)
Min. term frequency:	50
Min. percentage of documents with a term:	0.1%
Max. percentage of documents with a term:	40%

Preprocessing additional options:

- Dictionary/CDM vocabulary search
- Specific regex rules

Text representations:

- Bag-of-words
- TFIDF
- To be implemented:
  - *Topic Models (LDA)*
  - *Embeddings*
    - *Word (GloVe)*
    - *Document (GloVe aggregated, Doc2vec)*
  - *Transformers (BERT, BioBERT)*

Discussion: Information in Coded data vs Clinical text

- Depends on:
  - Database: EHR (lot of text) vs claims (mainly coded)
  - Problem settings with much text and few coded data:
    - Psychology/Depression
    - Family situations
    - Lifestyle
- If structured data is well-coded (high quality), the clinical text will not provide additional information.

Tom M. Seinen, Jan A. Kors, Erik M. van Mulligen, Peter R. Rijnbeek

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.