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: N-gramic representation
- **Proof of concept study**
  - Database: Integrated Primary Care Information (IPC)
  - 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 n-grams, bag-of-words/TIFIDF (365d)
- **Experimental setup**

**RESULTS**

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC (IC)</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Structured</td>
<td>0.67 (0.56-0.79)</td>
<td>0.012</td>
</tr>
<tr>
<td>2a Text n=1</td>
<td>0.68 (0.56-0.97)</td>
<td>0.011</td>
</tr>
<tr>
<td>2b Text n=2</td>
<td>0.69 (0.58-0.80)</td>
<td>0.012</td>
</tr>
<tr>
<td>3a Struct. + Text n=1</td>
<td>0.80 (0.73-0.88)</td>
<td>0.022</td>
</tr>
<tr>
<td>3b Struct. + Text n=2</td>
<td>0.77 (0.69-0.83)</td>
<td>0.017</td>
</tr>
</tbody>
</table>

**NLP pipeline settings:**
- **Preprocessing:** Lowercase; digit and symbol removal
- **Tokenization:** Word tokenization (strings)
- **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
- TF-IDF

**To be implemented:**
- Topic Models (LDA)
- Embeddings
  - Word (GloVe)
  - Document (GloVe-averaged, Doc2vec)
  - Transformers (BERT, BioBERT)

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

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