First results

Extraction of signs and symptoms from Dutch primary care electronic health records notes: a comparison of NLP approaches

Matthew Scheeres*, Isa Spiero*, Merijn Rijk, Roderick Venekamp, Artuur Leeuwenberg

UMC Utrecht

Julius Center for Health Sciences and Primary Care

Setting

Hospitalization/mortality prediction model in General Practice (GP)

- Cohort: primary care patients with Lower Respiratory Tract Infection (LRTI)

- Outcome: hospitalization or mortality within 30 days

- Possible predictors:
  - patient-reported symptoms
  - GP-reported signs
  - vital and laboratory measurements
  - LRTI diagnosis codes, medical history, medication, demographics

➢ **Extract these data from primary care EHRs**

UMC Utrecht

Problem

- **Problem**: some predictors only available in free text notes

    - Need to screen & extract from many EHRs to have a sufficient number of patients/data for prediction modelling

    - Possible solution: **Dutch Large Language Models (LLMs)** to extract

UMC Utrecht

Research aim

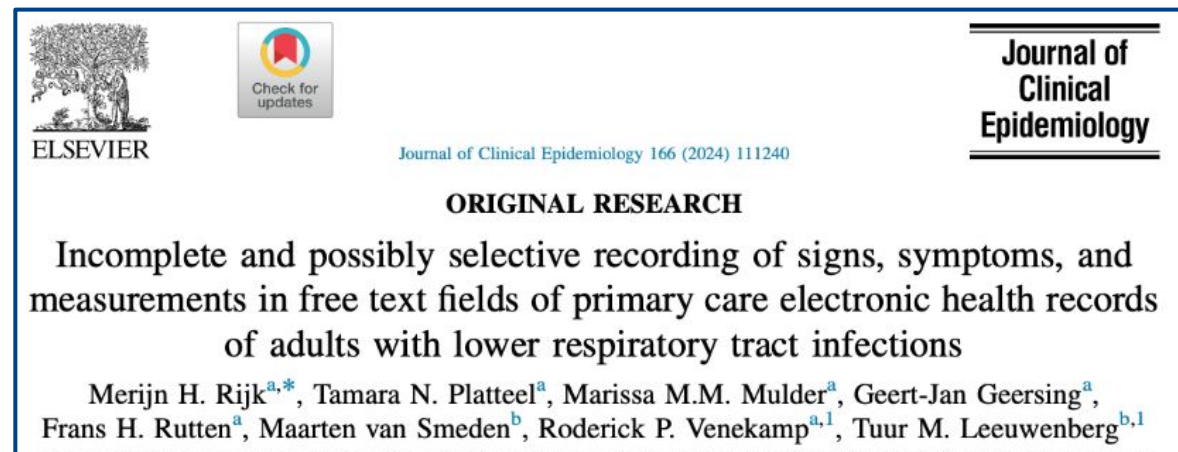(1) Comparing Dutch LLMs to extract predictors from general practice EHRs for prediction modelling

and

(2) Determining required training sample size for sufficient extraction performance

UMC Utrecht
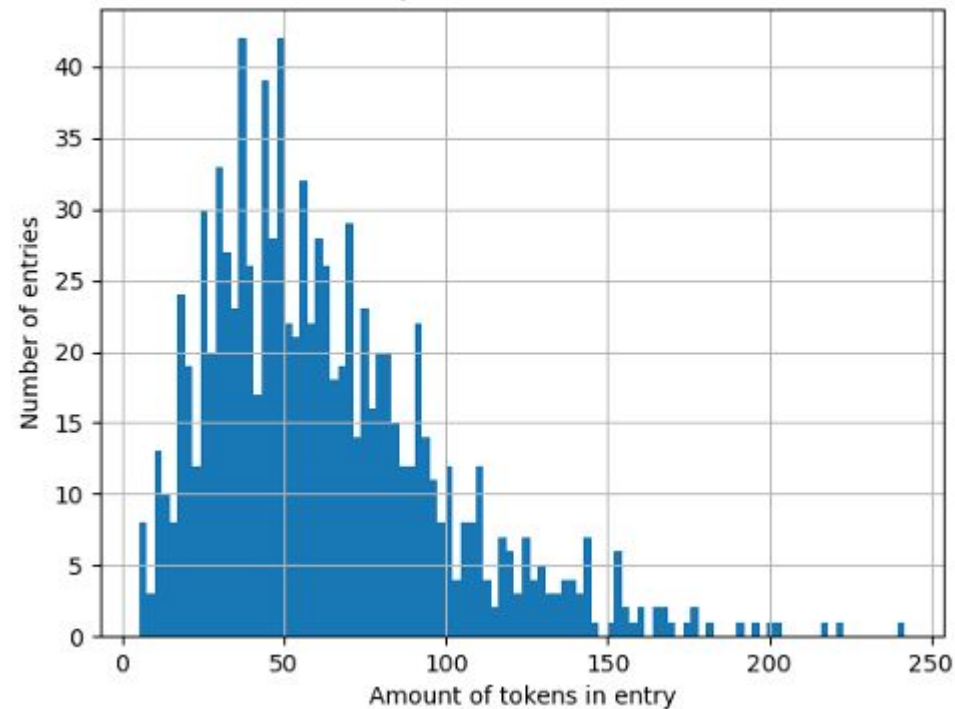
Methods

Dataset

- EHR data source: Julius General Practitioners' Network in the region of Utrecht

- Clinical note type: 'SOEP' (subjective, objective, evaluation, plan) reports

- Gold standard: 1000 manually labeled reports for LRTI signs and symptoms (Rijk *et al.,* 2024)

Journal of Clinical Epidemiology 166 (2024) 111240

ORIGINAL RESEARCH

Incomplete and possibly selective recording of signs, symptoms, and measurements in free text fields of primary care electronic health records of adults with lower respiratory tract infections

Merijn H. Rijk[a,*], Tamara N. Platteel[a], Marissa M.M. Mulder[a], Geert-Jan Geersing[a], Frans H. Rutten[a], Maarten van Smeden[b], Roderick P. Venekamp[a,1], Tuur M. Leeuwenberg[b,1]

UMC Utrecht

Methods

Dataset

- Number of tokens per SO(EP) report: mostly between 5-100 (mean: 63)

Methods

LRTI-related symptoms (predictors) to be extracted by the LLMs

- Cough ('hoesten')
- Fever ('koorts')
- Shortness of breath/dyspnea ('kortademigheid')

For each symptom the LLM performs multiclass classification

Possible classes:
(a) symptom 'recorded as positive'
(b) symptom 'recorded as negative'
(c) symptom 'not recorded'

UMC Utrecht

Methods

Fictional example of sample in dataset

Patnr	start_epi	start_icpc	SOEPcode	*Koors*	*Hoesten*	*Kortademigheid*	DEDUCE_omschrijving
100020	2018-02-27	R90	SO	1	2	0	*"Pat heeft sterke verhoging, ademhaling goed"*

UMC Utrecht

Methods

Overview of dataset

	Pneumonia	Acute bronchitis	Overall	Recorded as positive	Recorded as negative	Not recorded
Patient reported						
Cough	71.9	82.6	76.6	75.4 (98.4)	1.2 (1.6)	23.4
Fever	57.0	50.3	54.1	31.2 (57.7)	23.4 (43.3)	45.9
Shortness of breath	54.4	51.7	53.2	37.7 (70.9)	15.5 (29.1)	46.8

Extracted symptoms

- Often not recorded

Methods

	Pneumonia	Acute bronchitis	Overall	Recorded as positive	Recorded as negative	Not recorded
Patient reported						
Cough	71.9	82.6	76.6	75.4 (98.4)	1.2 (1.6)	23.4
Fever	57.0	50.3	54.1	31.2 (57.7)	23.4 (43.3)	45.9
Shortness of breath	54.4	51.7	53.2	37.7 (70.9)	15.5 (29.1)	46.8

Extracted symptoms

- Often not recorded

- If recorded: imbalanced (e.g. 'Cough')

Methods

	Pneumonia	Acute bronchitis	Overall	Recorded as positive	Recorded as negative	Not recorded
Patient reported						
Cough	71.9	82.6	76.6	75.4 (98.4)	1.2 (1.6)	23.4
Fever	57.0	50.3	54.1	31.2 (57.7)	23.4 (43.3)	45.9
Shortness of breath	54.4	51.7	53.2	37.7 (70.9)	15.5 (29.1)	46.8

Extracted symptoms

- Often not recorded

- If recorded: imbalanced (e.g. 'Cough')

Methods

Dutch LLMs that were compared

Direct classifiers

- MedRoBERTa.nl
- RobBERT

Prompt-based

- MedRoBERTa.nl
- RobBERT

UMC Utrecht

Methods

Dutch LLMs that were compared

Direct classifiers

- **MedRoBERTa.nl**
- RobBERT

Prompt-based

- MedRoBERTa.nl
- RobBERT

MedRoBERTa.nl (Verkijk *et al.*, 2021)

- RoBERTa-based
- Trained on 13GB Dutch EHR data
- 117 million parameters
- Relatively small: locally applicable

UMC Utrecht

Methods

Dutch LLMs that were compared

Direct classifiers

- MedRoBERTa.nl
- **RobBERT**

Prompt-based

- MedRoBERTa.nl
- RobBERT

RobBERT (Delobelle *et al.*, 2020)

- BERT (multilingual) based
- 117 million parameters
- Also relatively small

UMC Utrecht

Delobelle, P., Winters, T., & Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based Language Model, Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online.

Methods

Dutch LLMs that were compared

Direct classifiers

- MedRoBERTa.nl
- RobBERT

Prompt-based

- MedRoBERTa.nl
- RobBERT

- Fine-tuned for sequence-to-sequence generation
- HealthCareMagic-100k dataset translated to Dutch with Google Translate

UMC Utrecht

Methods

Training settings

- **Direct classifiers**

  - Training sample size variations: 1, 3, 6, 12, 25, 50, 100, 200, 400 and 800

  - 5-fold cross-validation

  - Trained for 5 epochs

- **Prompt-based classifiers**

  - Prompt example variations: 1, 2, and 3

UMC Utrecht

Methods

Evaluation

For each setting

- Recall
- Precision
- F1-score (macro)

Model type	Model name	Target variable	Variations
Direct Classifiers	MedRoBERTa.nl	Fever	1-800 samples
		Cough	1-800 samples
		Shortness of breath	1-800 samples
	RobBERT	Fever	1-800 samples
		Cough	1-800 samples
		Shortness of breath	1-800 samples
Prompt-Based Classifiers	MedRoBERTa.nl	Fever	1-3 samples
		Cough	1-3 samples
		Shortness of breath	1-3 samples
	RoBERT	Fever	1-3 samples
		Cough	1-3 samples
		Shortness of breath	1-3 samples

UMC Utrecht

Results

Direct classifiers:

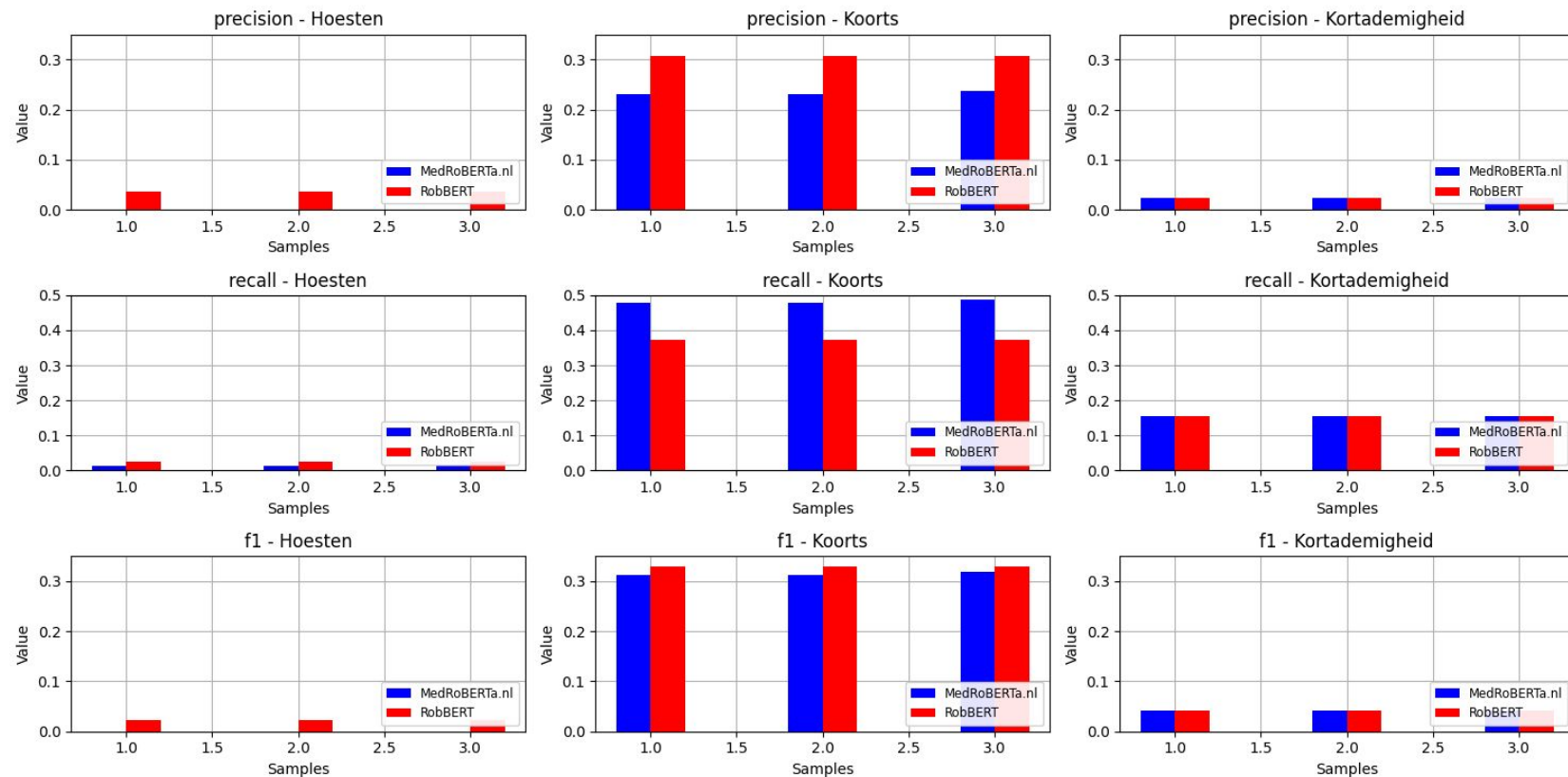Depends on symptom, and sample size

- RobBERT
  - Often better

- MedRoBERTa
  - Best for largest sample size
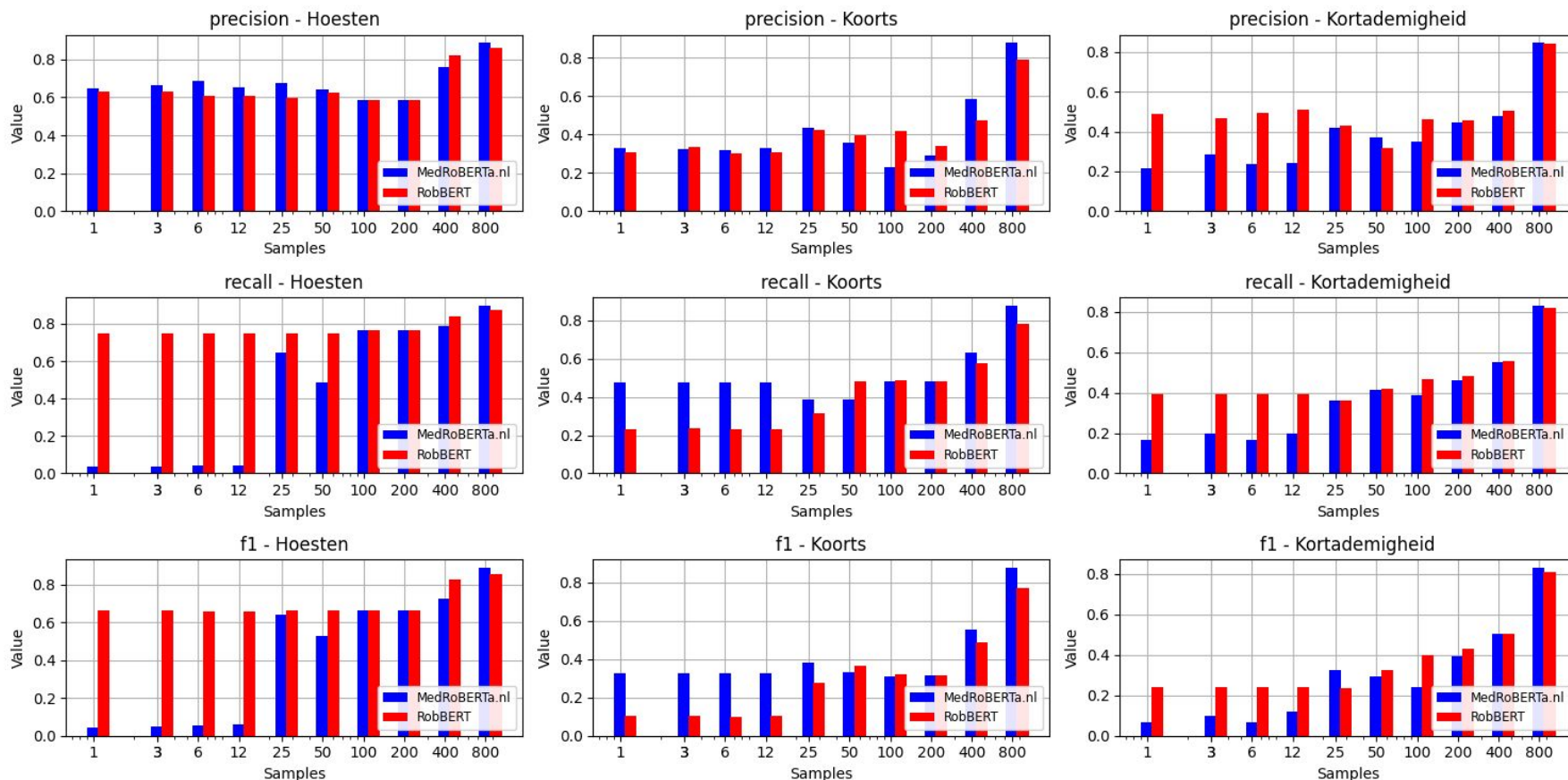
Results

Prompt-based:

- Similar model performance

- No increased performance with more examples in the prompt

Discussion/Conclusion

Best performing model

- Direct classifiers better than prompt based
- MedRoBERTa.nl slightly better than RobBERT (F1-score up to 0.88)

Training set size

- Direct: 800 or more needed for training for sufficient performance
- Prompt: No increased performance with more examples in the prompt

UMC Utrecht

Discussion/Conclusion

Finding optimal performance limited by

- Local applicability of LLMs
- Availability of pre-trained Dutch LLMs
- Lack of annotated EHR data

Future research

- Other LRTI-related signs and symptoms: lower prevalence
- Other prompt designs: 'optimal' prompt by prompt-tuning
- Added value of signs and symptoms in prediction of hospital admission or mortality

UMC Utrecht

Questions?

Matthew Scheeres*, Isa Spiero*, Merijn Rijk, Roderick Venekamp, Artuur Leeuwenberg

UMC Utrecht

Julius Center for Health Sciences and Primary Care

Funding

RAISE (NWA.1418.22.008)

ELEMENT (08391052110003)