BC Primary Health Care RESEARCH NETWORK
www.spor-bcphcrn.ca @BC_PHCRN

Using Machine Learning to Improve De-identification Accuracy in Primary Care Electronic Medical Record Data

Yale Wang[1], Andy Gibb[1], Sabrina Wong[1,2], Rubee Dev[1,2], Anita Kapanen[3], Peter S. Loewen[3], Peter J. Zed[3]
[1] Centre for Health Services and Policy Research, UBC; [2] School of Nursing, UBC; [3] Faculty of Pharmaceutical Sciences, UBC

Background

- The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) extracts data from consenting providers' electronic medical record (EMR) systems across Canada. BC-CPCSSN is the BC node of the project, based at UBC.
- BC-CPCSSN collaborated with the UBC Faculty of Pharmaceutical Sciences' Pharmacists in Primary Care Network program to analyze over 100,000 pharmacists' notes.
- Personal identifiers must be removed from the data before analysis.
- CPCSSN currently de-identifies data by matching words in EMR records with a list of known names from past records and public censuses.
- More text data means uncommon names become more likely.
- This study aims to improve de-identification accuracy by employing a novel machine learning approach.

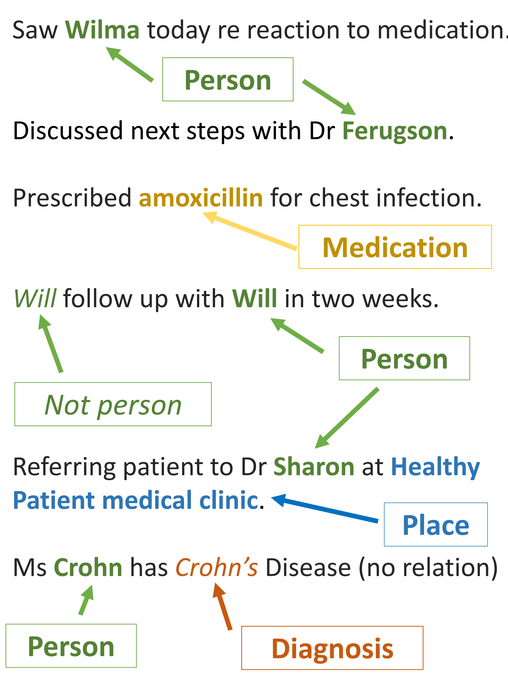
Methods

- A sample of 500 pharmacists' notes were manually labeled as a reference selection; 73 names were identified.
- CPCSSN's existing de-identification method and five additional pre-trained machine learning models were chosen from open-source libraries.

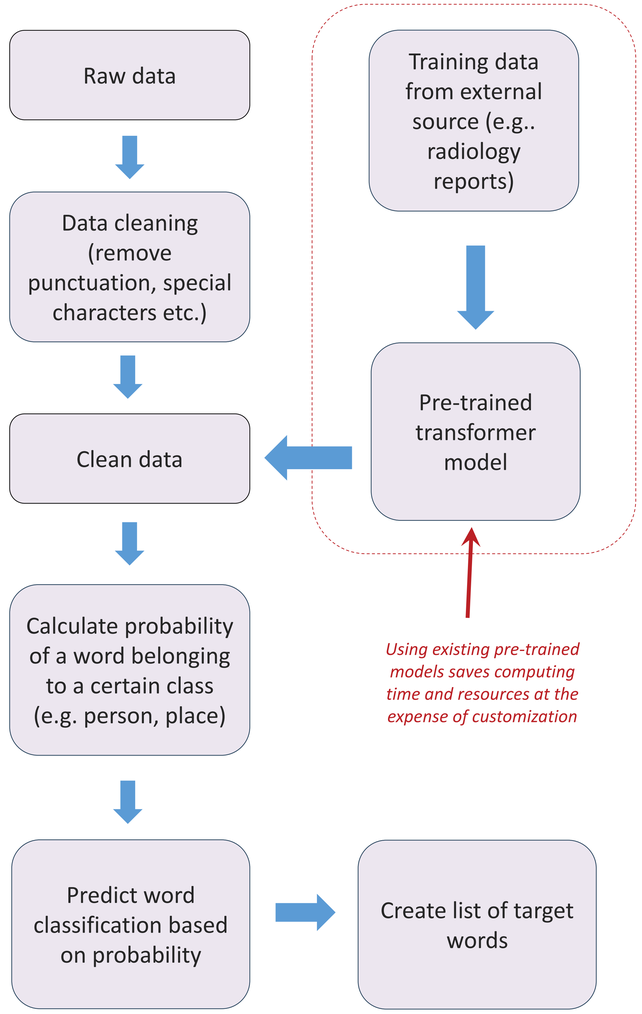
Problem

How to identify unknown names from thousands of lines of text?

Solution: Named-entity recognition



Implementation of pre-trained machine-learning models



Results for each model

	CPCSSN	Spacy English Model Only	Spacy English Model + Spacy Medical Model	English Model + Medical Term List	Stanford Deidentification model
Precision	**93.8 %**	41.2 %	56.2 %	56.2 %	92.0 %
Recall	83.5 %	**100 %**	68.5 %	97.3 %	94.5 %
F1	88.4 %	58.4 %	61.7 %	71.2 %	**93.2 %**

Precision *(aka positive predictive value) measures the proportion of true positive cases to all positive cases = true positives / (true positives + false positives)*
Recall *(aka sensitivity) measures the ratio of the number of true positive cases to all correct cases = true positives / (true positives + false negatives)*
F1 *score is the harmonic mean of precision and recall = 2 × precision × recall / (precision + recall)*

- These models were applied in turn to the same set of notes.
- Names were identified by model from context ("named-entity recognition").
- Model performance was assessed using F1 scores, where a higher score indicates more accurate name identification.
- The highest-scoring model was applied to full dataset.

Results

- Existing de-identification method achieved an F1 score of 88 %.
- The Hugging Face model, trained by Stanford University, outperformed others with an F1 score of 93 %.
- Other selected models, such as the Spacy Natural Language Processing Model, and their combinations yielded F1 scores considerably lower than the existing method.
- Identified misspellings.
- However, word list included many more words than names alone, requiring human intervention to finalize list.

Conclusions

- Machine learning models show clear potential for enhancing the de-identification of EMR records.
- These models also aid in mitigating scalability issues associated with searching larger datasets for names.
- Limiting factor is human intervention still required to remove non-names.
- Further research on larger datasets is warranted to demonstrate scalability effectively.

Strategy for Patient-Oriented Research SPOR Putting Patients First
Canadian Institutes of Health Research / Instituts de recherche en santé du Canada
Canada