

International Journal of Population Data Science

Journal Website: www.ijpds.org



Using natural language processing on free-text notes found in family medicine EMRs to support healthcare system research

Liisa Jaakkimainen¹, Elizabeth Monchesky², Shrey Virmani², Vanessa Lu², and Ali Moinshaghghi²

¹ICES, Toronto, Canada; University of Toronto, Toronto, Canada

²ICES, Toronto, Canada

Objectives

Family physician (FP) electronic medical records (EMRs) have been linked to health administrative data to calculate primary care wait times for cancer patients. Since FP EMRs contain non-structured free-text notes, we assessed the feasibility of using natural language processing (NLP) on FP EMR notes improve the efficiency of identifying the index date for abnormal results.

Approach

We used a convenience sample of 385 FPs across Ontario, Canada who consented to have their EMR linked to health administrative data. In prior studies we had a trained abstractor review the entire EMR to identify the index date for patients having colon, endometrial, bladder or melanoma cancers which represented the first indication of an abnormal symptom, sign or radiological result found in their FPs EMR. Transformer models were trained using progress notes, consults and referrals using the labels identified by the abstractor.

Results

ClinicalBERT was best model for bladder (precision 0.73, recall 0.94, F1 0.82 and accuracy 0.96), colon (precision 0.72, recall 0.74, F1 0.73 and accuracy 0.94) and melanoma cancers (precision 0.87, recall 0.93, F1 0.90 and accuracy 0.98). Longformer was the best model for endometrial cancer (precision 0.68, recall 0.85, F1 0.75 and accuracy 0.95).

Conclusion

The NLP models which used FP EMR data had reasonable precision and recall in identifying cancer index dates. Implications: The use of NLP models on FP EMR free-text notes can improve the efficiency of using FP EMR linked to administrative data to support healthcare system tracking of primary care wait times.

