Comparing Three Methods of Extracting COVID-19 Related Symptoms from EHR Data in a Large Healthcare System

Hannah A. Burkhardt[1], Nicholas Dobbins[1], Kevin Lybarger[1], Brenda Mollis[1], Margaret Au[1], Kris Pui Kwan Ma[1], Meliha Yetisgen[1], Angad Singh[1], Matthew Thompson[1], Kari A. Stephens[1]

[1]University of Washington, Seattle, WA, USA

COVID-19 symptom data in the electronic health record (EHR) is a key resource to understand and predict COVID-19's disease progression and to support pandemic recovery

Introduction

The COVID-19 pandemic has claimed over 220,000 lives in the United States[1]. A promising resource for discovery in COVID-19's symptom progress is data documented in electronic health record (EHR) systems as part of clinical care. Such data are stored in disparate locations within the EHR, requiring multiple extraction methods. We compared the symptom detection rates of three extraction methods to assess the comparative utility of each source of COVID-19 related symptoms within the EHR.

Methods

- Symptoms were extracted from EHR data for all patients who were tested for SARS CoV-2 through May 31, 2020 from a single large healthcare system in the state of Washington.
- Three methods of extraction used:
  - Extraction of ICD-10 codes
  - Regular expression matching of clinical notes using a template developed for standard use across the health system
  - A Natural Language Processing (NLP) pipeline[2,3] applied to clinical notes
- We conducted descriptive statistics on the unique and overlapping symptoms detected by each extraction method for symptoms that were documented in the EHR within 10 days prior to SARS CoV-2 PCR lab test.

Results

- SARS CoV-2 PCR tests were conducted across 24,775 unique patients, who were given 32,924 total tests between February 29 and May 31, 2020.
- The study cohort was refined to 14,159 patients who had a test and an encounter (excluding "Orders Only") with provider during the study period.
- COVID-19 related symptoms were extracted at differential rates across sources within the EHR (see Fig.1).
- On average, tested patients had 2.5 (SD 2.7) symptoms documented within 10 days prior to a SARS CoV-2 PCR test. However, 40% of tests had no associated symptoms identified.
- NLP detected the most symptoms of all extraction methods, 25,433 (91.9%) of symptoms. 17,904 (64.7%) of symptoms were detected *only* by NLP.
- The ICD data source added 1,969 (7.1%) symptoms that were not already captured by NLP.
- Parsing of notes using regular expression extraction from a known structure added 276 (1.0%) more symptoms.

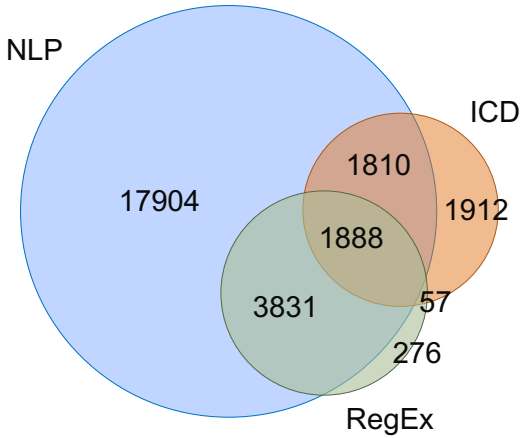

Figure 1. COVID-19 related symptom totals and overlap between extraction methods.

Discussion

- All three extraction methods contributed to COVID-19 symptom detection, with NLP detecting the large majority of symptoms and ICD coded data detecting the least number of symptoms.
- A standardized note template containing a discrete checklist of COVID-19 related symptoms led to simple and highly accurate text parsing. However, the template was used infrequently, and NLP extraction was able to parse most of the template-derived symptoms.
- Given NLP methods resulted in the highest extraction rate of COVID-19 related symptoms, using only methods such as regular expression extraction and structured data extraction of ICD codes may miss a significant amount of symptom data.

Acknowledgements

This project was supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1 TR002319.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020;**20**:533–4. doi:10.1016/S1473-3099(20)30120-1
2. uw-bionlp/uwbionlp-parser. https://github.com/uw-bionlp/uwbionlp-parser (accessed 24 Aug 2020).
3. Yetisgen M, Vanderwende L, Black T, *et al.* A New Way of Representing Clinical Reports for Rapid Phenotyping. In: *Proceedings of AMIA 2016 Joint Summits on Translational Science*. San Francisco: 2016.