Extracting COVID-19 Related Symptoms from EHR Data: A Comparison of Three Methods

Hannah A Burkhardt[1]

haalbu@uw.edu

Nicholas Dobbins[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

Introduction

- COVID-19 has claimed >500,000 U.S. lives(Dong, Du, and Gardner 2020).
- **Electronic health record (EHR) data is a promising resource for COVID-19 symptom research**.
- Symptom data are stored in multiple 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 EHR-source of COVID-19 related symptoms.

Methods

- Associated symptoms were extracted from EHR data for all SARS CoV-2 tests through May 31, 2020 conducted by a single large healthcare system in WA. Three methods were used:
  1. ICD-10 codes (structured symptom & diagnosis data documented for medical billing),
  2. regular expression matching of notes utilizing the health system's **COVID-19 screening note template**, and
  3. a previously reported and evaluated Natural Language Processing (NLP) pipeline (Yetisgen et al. 2016; uw-bionlp n.d.) applied to **clinical notes**.
- ICD codes, NLP, and pattern parsing outputs were matched to one (or none) of 11 symptoms.
- Presence or absence of each symptom in the **10 days prior to SARS CoV-2 PCR lab test** was determined for each of the 3 extraction methods
- To validate NLP performance, automatically extracted symptoms were compared to manual annotations in a small sample of notes.

Results

- **32,924 COVID-19 tests** were conducted for **25,115 unique patients** between February 29 and May 31, 2020 (5.9% positive).
- The 3 sources yielded COVID-19 related symptoms at differential rates.
- On average, **tested patients had 1.1 (SD 1.9) symptoms** documented within 10 days before a SARS CoV-2 PCR test, with myalgia (21.9%) being the most common. **65.4% of tests had no associated symptoms** identified (Figure 1).

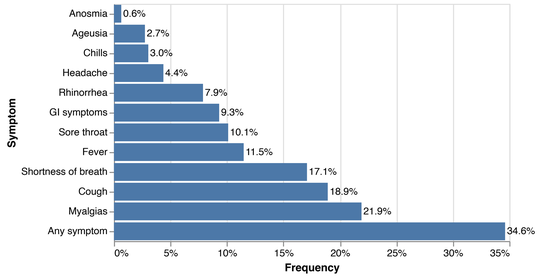

Figure 1: Percentage of tests where the patient had the symptom recorded in the prior 10 days.

Clinical notes are a key resource for understanding COVID-19 symptoms, to predict COVID-19 disease progression, and to support pandemic recovery.

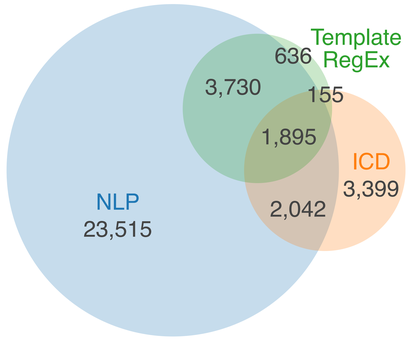

Figure 2: COVID-19 related symptom totals and overlap between extraction methods.

- **NLP detected the most symptoms** (88.2% of all symptoms). 66.5% were detected only by NLP (Figure 2, Figure 3).
- The ICD data source added 3,554 (10.0%) symptoms that were not already captured by NLP, and the regular expression parsing added 636 (1.8%) more symptoms (Figure 3).
- In a small sample of 10 manually annotated notes, NLP demonstrated an average sensitivity of 79% and an average specificity of 77%.

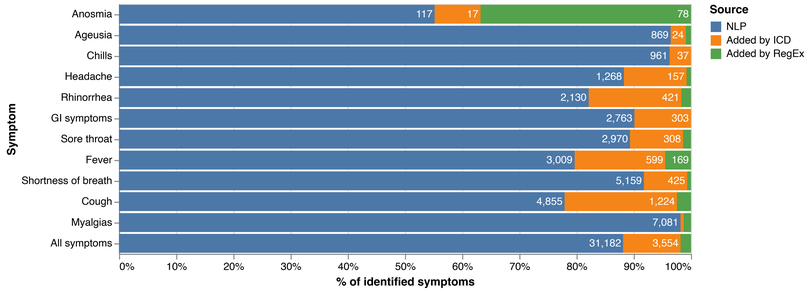

Figure 3: Number of new symptoms indentified by adding sources in the shown order.

Discussion & Conclusion

- All three extraction methods added unique symptoms. **NLP detected the large majority of symptoms**. Template parsing detected the least.
- Parsing the standardized COVID-19 screening template was simple and accurate; however, the template was used infrequently, and NLP also found most of the template-derived symptoms.
- NLP captured more symptoms than ICD codes, because clinical narrative may be more detailed and capture information peripheral to the chief complaint. However, more false positives should be expected from NLP than structured data.
- **Structured data alone may miss a significant amount of symptom data**.

Acknowledgements

This work was supported by NLM Biomedical and Health Informatics Training Grant 5T15LM007442-19, the Gordon and Betty Moore Foundation, and by the National Center For Advancing Translational Sciences of the National Institutes of Health (UL1 TR002319).

References

Dong, Ensheng, Hongru Du, and Lauren Gardner. 2020. "An interactive web-based dashboard to track COVID-19 in real time." *The Lancet Infectious Diseases* 20 (5): 533–34. https://doi.org/10.1016/S1473-3099(20)30120-1.

uw-bionlp. n.d. "uw-bionlp/uwbionlp-parser: A container-based CLI for executing various NLP algorithms on documents, developed by @uw-bionlp." Accessed August 24, 2020. https://github.com/uw-bionlp/uwbionlp-parser.

Yetisgen, Meliha, Lucy Vanderwende, Tony Black, Sean Mooney, and Peter Tarczy-Hornoch. 2016. "A New Way of Representing Clinical Reports for Rapid Phenotyping." In *Proceedings of Amia 2016 Joint Summits on Translational Science*. San Francisco.