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

Hannah A. Burkhardt[1], 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, USA

Introduction

The COVID-19 pandemic has claimed over 310,000 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 were used: 1) extraction of ICD-10 codes, which reflected symptoms and diagnoses documented for medical billing, 2) regular expression matching of clinical notes using a COVID-19 note template developed for standard use across the health system, and 3) a previously reported and evaluated Natural Language Processing (NLP) pipeline[2,3] applied to clinical notes. Patients were considered to either have or not have each of 11 different symptoms (fever, cough, shortness of breath, sore throat, rhinorrhea, headache, GI symptoms, general aches and pains (myalgia), anosmia, ageusia, and chills) by each of the 3 methods if they were documented in the EHR in the 10 days prior to SARS CoV-2 PCR lab test. ICD codes and NLP and pattern parsing outputs were matched to one of the 11 symptoms. We obtained descriptive statistics on the unique and overlapping symptoms detected by each of these extraction methods. A small sample of notes was manually annotated for symptom presence by the authors and compared to automatically extracted symptoms to validate NLP performance.

Results

SARS CoV-2 PCR tests were conducted across 25,115 unique patients, who were given 32,924 total tests between February 29 and May 31, 2020. COVID-19 related symptoms were extracted at differential rates across sources within the EHR (see Figure 1). On average, tested patients had 1.1 (SD 1.9) symptoms documented within 10 days before a SARS CoV-2 PCR test, with cough (24%), myalgia (23%), and fever (20%) being the most common. However, 65% of tests had no associated symptoms identified. NLP detected the most symptoms of all the extraction methods, namely 88.2% of all symptoms, and 66.5% were detected only by NLP. The ICD data source added 3,554 (10.0%) symptoms that were not already captured by NLP, and the parsing of notes using regular expression extraction from a known structure added 636 (1.8 %) more symptoms. In a small sample of 10 manually annotated notes, NLP demonstrated an average sensitivity of 79% and an average specificity of 77%.

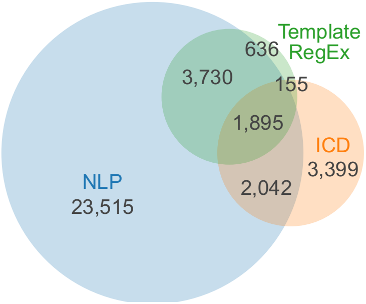

Figure 1. COVID-19 related symptom totals and overlap between extraction methods.

Discussion & Conclusion

All three extraction methods contributed to COVID-19 symptom detection, with NLP detecting the large majority of symptoms and template parsing 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. ICD codes directly provide discrete symptom data; however, NLP captured more symptoms than ICD codes, possibly because clinical narrative tends to be more detailed and captures information peripheral to the chief complaint. Given NLP methods resulted in the highest extraction rate of COVID-19 related symptoms, using only methods such as note template parsing and structured data extraction of ICD codes may miss a significant amount of symptom data.

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.