Identification of Post-acute Sequelae of COVID-19 from Electronic Health Records

Victor M Castro, MS[1,2], Vivian S Gainer, MS[1], Justin Manjourides, PhD[2,3], Roy H Perlis, MD[4], Shawn N Murphy, MD, PhD[1,4]

[1]Mass General Brigham, Somerville, MA; [2]Northeastern University, Boston, MA; [3]OHDSI Center at The Roux Institute, Northeastern University, Portland, ME; [4]Massachusetts General Hospital, Boston, MA

Objective

The goal of this work is to build a generalizable and portable library of symptoms defined from EHR data linked to common vocabularies for use in observational studies of COVID-19 and post-acute sequalae of COVID (PASC).

Background

- Identifying signs, symptoms and other ill-defined conditions is notoriously difficult in electronic health record (EHR) and claims data.
- While billing codes, problem list and other structured fields exist, they typically suffer from low sensitivity and variable use across data sources.
- Accurate identification of symptomology has become increasingly important during the COVID-19 pandemic.

Methods

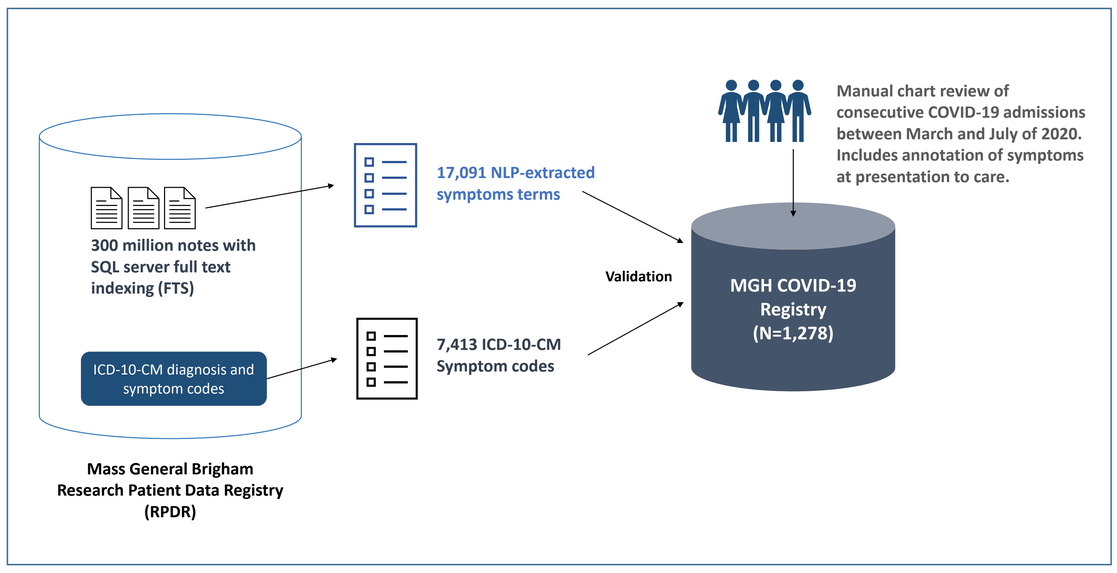

Figure 1. Study schematic

Symptoms

We identified a set of 16 COVID and PASC-related symptoms that were reported in the literature and had available validation data.

ICD-10-CM Symptom Codes

For each symptom, we identified appropriate ICD-10-CM diagnosis codes in the symptom chapter (ICD-10 R*) and other related disease codes.

Natural Language Processing (NLP)

We also developed a natural language processing (NLP) approach to identification of symptoms by manually and iteratively curating token lists derived from survey and other literature sources. These token lists were combined with rules for negation and family history exclusion and implemented as structured query language (SQL) full text index queries. All clinical notes during a COVID-19 admission at MGH were then loaded into a SQL server database with full-text indexing enabled to allow rapid evaluation and iteration of the NLP rules. Full-text SQL querying is implemented in most major relational databases allowing portability and implementation in both i2b2 and OMOP common data models.

Gold Standard

The developed ICD-10 and NLP rules are evaluated against a 'gold standard' manually annotated registry of consecutive Massachusetts General Hospital (MGH) COVID-19 admissions between March and July 2020[1].

Download symptom definitions, code and additional results at:
https://github.com/vcastro/ehr_symptom_validation

Results

Table 1. Sociodemographic characteristics of the symptom validation dataset.

	COVID-19 Hospitalized March-July 2020 N = 1,278[1]		COVID-19 Hospitalized March-July 2020 N = 1,278[1]
Age (years)	60 (46, 73)	**Symptom**	
Gender		Congestion	138 (11%)
Female	542 (42%)	Headache	230 (18%)
Male	736 (58%)	Cough	831 (65%)
Race		Sore throat	212 (17%)
Asian	49 (3.8%)	Fatigue	354 (28%)
Black or African American	135 (11%)	Dyspnea (SOB)	702 (55%)
Hispanic	460 (36%)	Anorexia	195 (15%)
Other	32 (2.5%)	Nausea/Vomiting	249 (19%)
Unknown	110 (8.6%)	Diarrhea	273 (21%)
White	492 (38%)	Myalgia or arthralgia	474 (37%)
Pregnant	27 (2.1%)	Fever	845 (66%)
Housing status		Chills	257 (20%)
Homeless	33 (2.6%)	Rhinorrhea	112 (8.8%)
Other	64 (5.0%)	Anosmia	133 (10%)
Private home/apartment	959 (75%)	Dysgeusia	52 (4.1%)
SNF/Rehab/Assisted Living	222 (17%)	Altered mental status	126 (9.9%)

[1] n (%); Median (IQR); Range — [1] n (%)

Figure 2 below illustrates the specificity, sensitivity, F1 score and precision (PPV) of the ICD, NLP and combined ICD and NLP methods for identifying symptoms from EHRs. The NLP approach increased sensitivity substantially when compared to ICD codes alone, but generally ICD-10 codes were more precise. Improvements in NLP methods may increase overall accuracy.

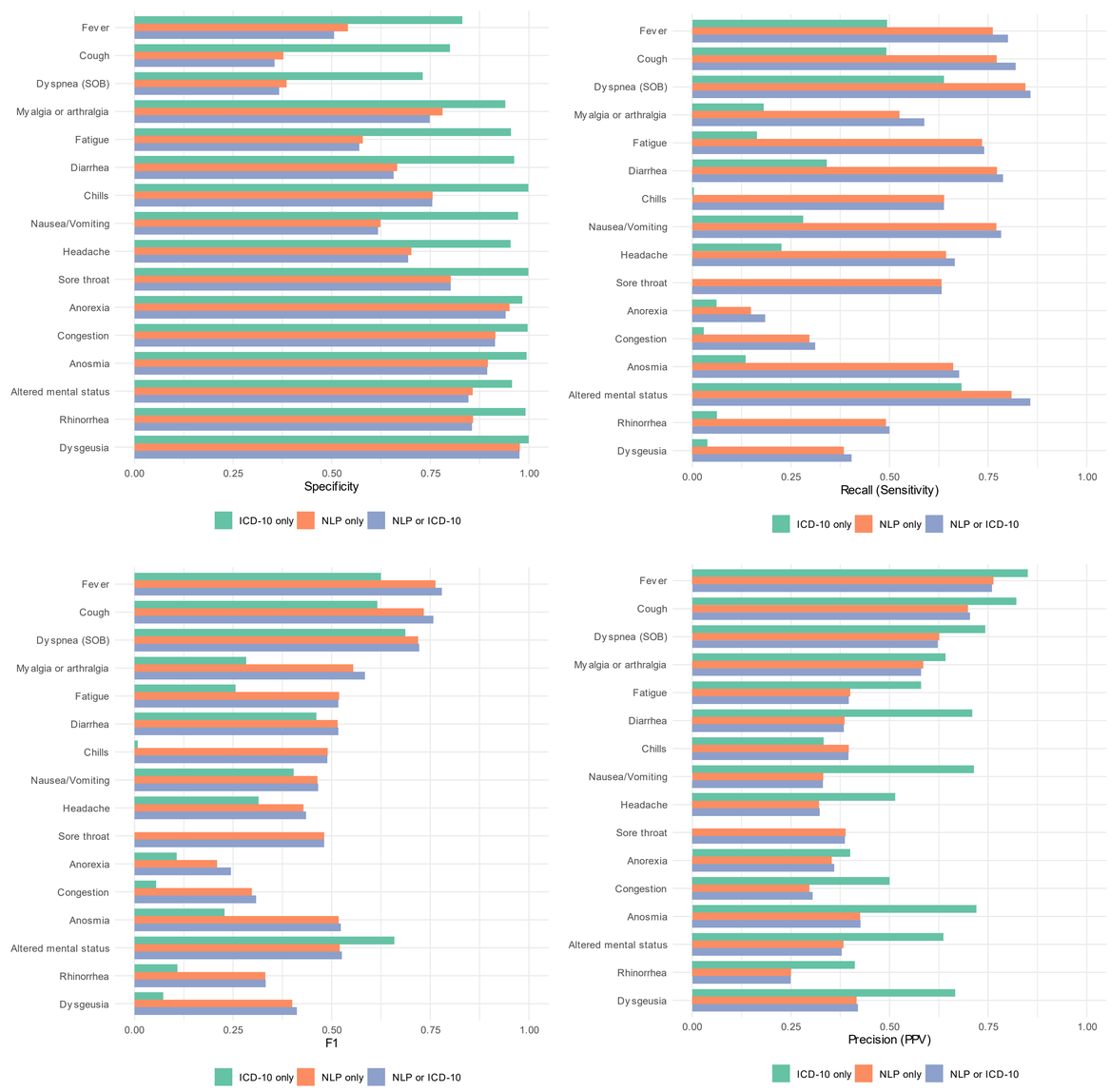

Figure 2. Performance metrics of ICD-10, NLP and combined methods for 16 COVID-19/PASC symptoms

Conclusion

In this work we have begun to develop and assess EHR symptom definitions to be used in observational and population health studies of COVID-19 and PASC. We use manually collected cohort study data to validate these EHR definitions so they can be accurately applied to larger populations available in EHR databases. Future work will improve NLP methods and assess performance of the definitions in under-represented populations.

References

1. Bassett IV, Triant VA, Bunda BA, Selvaggi CA, Shinnick DJ, et al. (2020) Massachusetts General Hospital COVID-19 registry reveals two distinct populations of hospitalized patients by race and ethnicity. PLOS ONE 15(12): e0244270. https://doi.org/10.1371/journal.pone.0244270