MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231087

A Symptom-Based Natural Language Processing Surveillance Pipeline for Post-COVID-19 Patients

Greg M. SILVERMAN[a,¶], Geetanjali RAJAMANI[b,¶], Nicholas E. INGRAHAM[c], James K. GLOVER[a], Himanshu S. SAHOO[d], Michael USHER[c], Rui ZHANG[a], Farha IKRAMUDDIN[e], Tanya E. MELNIK[c], Genevieve B. MELTON[a] and Christopher J. TIGNANELLI[a,1]

[a] *Department of Surgery, University of Minnesota, Minneapolis, MN, USA*
[b] *Medical School, University of Minnesota, Minneapolis, MN, USA*
[c] *Department of Medicine, University of Minnesota, Minneapolis, MN, USA*
[d] *Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA*
[e] *Department of Rehabilitation Medicine, University of Minnesota, Minneapolis, MN, USA*

Abstract. Post-acute sequelae of SARS CoV-2 (PASC) are a group of conditions in which patients previously infected with COVID-19 experience symptoms weeks/months post-infection. PASC has substantial societal burden, including increased healthcare costs and disabilities. This study presents a natural language processing (NLP) based pipeline for identification of PASC symptoms and demonstrates its ability to estimate the proportion of suspected PASC cases. A manual case review to obtain this estimate indicated our sample incidence of PASC (13%) was representative of the estimated population proportion (95% CI: 19±6.22%). However, the high number of cases classified as indeterminate demonstrates the challenges in classifying PASC even among experienced clinicians. Lastly, this study developed a dashboard to display views of aggregated PASC symptoms and measured its utility using the System Usability Scale. Overall comments related to the dashboard's potential were positive. This pipeline is crucial for monitoring post-COVID-19 patients with potential for use in clinical settings.

Keywords. PASC, NLP, disease surveillance

1. Introduction

Post-acute sequelae of SARS CoV-2 (PASC), or "long COVID," is an umbrella diagnosis characterizing a group of conditions in which patients previously infected with COVID-19 experience long-term health problems persisting weeks to months post-infection [1,2]. Current studies estimate the prevalence of PASC in the population to be anywhere between 19 to 80% [1,3,4]. Symptoms are often systemic, and commonly include fatigue, dyspnea, cough, cognitive difficulties, depression, gastrointestinal dysfunction, among others [1,5]. While a formal definition of PASC is lacking, it is generally accepted that

[1]Corresponding Author: Christopher J. Tignanelli, email: ctignane@umn.edu.
[¶]Equally contributing authors

PASC is defined by the continuation or development of one or more symptoms for weeks to months following initial COVID-19 infection [1,3,6].

To the best of our knowledge, no system based on natural language processing (NLP) currently exists to identify "at risk" post-COVID-19 patients in need of PASC screening. A PASC disease surveillance system can be used to: (1) facilitate patient-level triage to "long COVID" treatment clinics [6], (2) monitor the burden of PASC, and (3) monitor treatment response. Surveillance is particularly important because the effects of PASC can be significantly reduced with early and aggressive multidisciplinary treatment [6].

1.1. Study Objectives and Motivation

Previously, our team demonstrated how natural language processing (NLP) based models could be used in acute COVID-19 and PASC symptomatology research [7–9]. In 2020, we developed a pipeline that automates the extraction of acute COVID-19 and PASC symptoms from clinical notes at scale across a network of 12 U.S. hospitals and 60 primary care clinics affiliated with the University of Minnesota (MHealth Fairview) as discussed in Silverman, *et al.* [7]. Our pipeline enables NLP-based PASC disease monitoring of a large cohort of patients diagnosed with COVID-19 through polymerase chain reaction **(**PCR).

This study has the following objectives:

1. Demonstrate through manual chart review by clinicians that our symptomatology-based pipeline can provide an estimate of patients likely to have PASC.
2. Create a PASC Surveillance Dashboard with longitudinal views of symptoms most likely due to acute COVID-19 and PASC.
3. Conduct a System Usability Scale (SUS) survey of a group of subject matter experts (SMEs) to assess the usability of the PASC Surveillance Dashboard.

2. Methods

Data for the COVID-19 surveillance pipeline (hereafter, COVID-19 pipeline), described in Silverman *et al.* [7], were provided by MHealth Fairview. There were 83,850 patients confirmed as positive for COVID-19 with a positive PCR test within MHealth Fairview between March 1, 2020 and October 26, 2022. NLP methods were used to identify 24,620 (29.36%) potential cases of PASC. 1026 patients from this cohort (1.22%) have been positively identified as having undergone a PASC screening, either through a confirmed encounter at the MHealth Fairview Adult Post-COVID-19 Clinic or through a given diagnosis of PASC as indicated by the ICD 10 code U09.9 in their problem list.

2.1. Data Processing & NLP Methods

Structured data available in the COVID-19 registry include patient demographics, labs, vitals taken at emergency department (ED) and outpatient (OP) visits, home medications taken for at least 3 months prior to the ED/OP visit, and comorbidities identified using ICD-10 codes. Unstructured data, including symptoms extracted from ED admission and OP clinical notes, were made available for each patient in the COVID-19 registry.

Methods described in Silverman, *et al.* [7], based on guidelines by the CDC and others [1,5,10], were used to define lexica for 23 symptoms related to acute COVID-19 and PASC (available here [11]). The language model rule-based lexical gazetteer described and validated in Sahoo, *et al.* [8] was used to extract symptoms from ED and OP notes. As of October 26, 2022, symptoms for 72,864 patients have been extracted for inclusion in the COVID-19 registry to track the emergence and progression of PASC symptoms [9].

2.2. PASC Surveillance Dashboard

The PASC Surveillance Dashboard was built using Tableau Desktop. Key visualizations were made available to users granted access via the University of Minnesota's Tableau Server. Postal code, existence of symptoms consistent with PASC, and PASC screening status (suspected versus confirmed) were aggregated and made available for prototype map views, while a deidentified set of data including symptoms from before and after the COVID-19 diagnosis date, sex, race, and age was made available for longitudinal views of symptom progression.. All data were sourced from the COVID-19 registry.

2.3. Evaluation Methods

A review of 153 randomly selected cases from the COVID-19 registry was conducted by three clinicians (F.I., N.I. and T.M.) with experience working with COVID-19 and PASC patients. PASC incidence was estimated by considering each patient's problem list, encounter diagnoses, dates, and presence or absence of symptoms in relation to the COVID-19 diagnosis date. Cases were categorized as "yes" for those most likely to have PASC, "no" for those most likely to not, and "indeterminate" for those requiring additional assessment.

Additionally, 5 SMEs in fields ranging from clinical practice to public health conducted a usability analysis of the PASC Surveillance Dashboard. Each participant completed the SUS survey, which asked them to anonymously rate ten statements about the dashboard's design, usability, and utility on a scale from "Strongly Disagree" to "Strongly Agree" [12].

3. Results

A patient at risk for developing PASC, and in need for further screening, was defined as someone with persistent symptoms thirty days or longer after a diagnosis of acute COVID-19. Given this definition, we estimated 29.36% of our patient population are suspected to have PASC as visualized in Figure 1(a-b). Evaluation of cases for classification of PASC for our patient sample yielded: 65 no PASC, 20 PASC, and 88 indeterminate. For the SUS survey there was a response rate of 71% (5 out of 7), with a mean composite score of 58, a maximum of 90, a minimum of 25, and a standard deviation of 27.97.

Figure 1a - Distribution of Patients with Symptoms Consistent with PASC

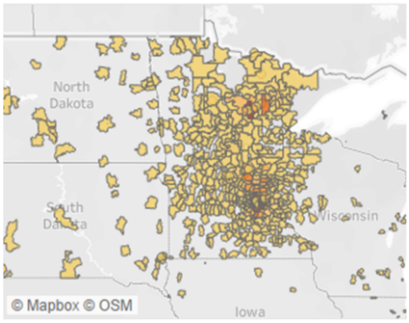
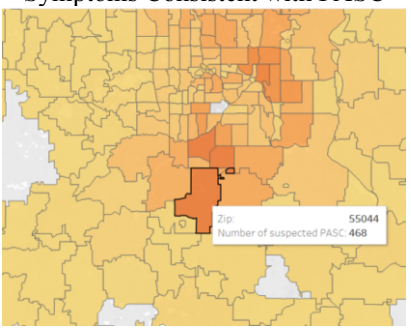


Figure 1b - High Concentration of Patients with Symptoms Consistent with PASC



4. Discussion

With 58% of cases classified as "indeterminate" the estimate of prevalence of PASC in our cohort is likely lower than the true incidence. This substantial proportion of indeterminate cases underscores the inherent difficulties encountered in accurately classifying PASC.

Evaluators found the PASC Surveillance Dashboard useful while also identifying areas for improvement. One useful set of views includes the geographic distribution of patients at risk who have suspected PASC symptoms shown in Figures 1(a–b). Detailed suggestions on how to improve the user experience were given, which are being integrated into a future release of the dashboard. Lastly, expansion of the lexicon for use in more precise classification of PASC, as per Wang, *et al.* is also being explored [14].

The main limitation of this study is the COVID-19 pipeline may overestimate patients in need of a PASC screening by flagging those with any symptoms detected in our lexicon, regardless of etiology. Conversely, it may also underestimate patients at risk for PASC if providers do not inquire about and/or document PASC-related symptoms. However, this pipeline serves as a valuable starting point for screening of patients at risk for PASC. Furthermore, the PASC Surveillance Dashboard can provide a method to identify those patients at risk for PASC, who may otherwise not receive proper care.

5. Conclusions

The PASC Surveillance Dashboard has potential for use in patient-level triage to "long COVID" treatment clinics by providing clinicians with tools needed to identify and monitor patients within a health system having suspected PASC. The COVID-19 pipeline infrastructure and visualization methods used for this study can easily be extended and applied to other outcomes and diseases based on signs and symptoms. This project represents an NLP-empowered translational informatics approach for taking models developed in a research lab setting directly into the clinic.

Acknowledgements

We acknowledge support from the University of Minnesota's Academic Investment Clinical Quality Program and the Center for Learning Health System Sciences and to express gratitude to those who graciously evaluated the PASC Surveillance Dashboard.

References

[1] CDC, Post-COVID Conditions, Centers for Disease Control and Prevention. 2022. https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html (accessed September 27, 2022).

[2] Cabrera Martimbianco AL, Pacheco RL, Bagattini ÂM, Riera R. Frequency, signs and symptoms, and criteria adopted for long COVID-19: A systematic review. Int J Clin Pract. 2021 Oct;75(10):e14357, doi:10.1111/ijcp.14357.

[3] Soriano JB, Murthy S, Marshall JC, Relan P, Diaz JV; WHO Clinical Case Definition Working Group on Post-COVID-19 Condition. A clinical case definition of post-COVID-19 condition by a Delphi consensus. Lancet Infect Dis. 2022 Apr;22(4):e102-7, doi:10.1016/S1473-3099(21)00703-9.

[4] Nearly One in Five American Adults Who Have Had COVID-19 Still Have "Long COVID," (2022). https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2022/20220622.htm (accessed November 26, 2022).

[5] Davis HE, Assaf GS, McCorkell L, Wei H, Low RJ, Re'em Y, Redfield S, Austin JP, Akrami A. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. EClinicalMedicine. 2021 Aug;38:101019, doi:10.1016/j.eclinm.2021.101019.

[6] Parker AM, Brigham E, Connolly B, McPeake J, Agranovich AV, Kenes MT, Casey K, Reynolds C, Schmidt KFR, Kim SY, Kaplin A, Sevin CM, Brodsky MB, Turnbull AE. Addressing the post-acute sequelae of SARS-CoV-2 infection: a multidisciplinary model of care. Lancet Respir Med. 2021 Nov;9(11):1328-41, doi:10.1016/S2213-2600(21)00385-4.

[7] Silverman GM, Sahoo HS, Ingraham NE, Lupei M, Puskarich MA, Usher M, Dries J, Finzel RL, Murray E, Sartori J, Simon G. Nlp methods for extraction of symptoms from unstructured data for use in prognostic covid-19 analytic models. J Artif Intell Res. 2021 Oct;72:429-74, doi:10.1613/jair.1.12631.

[8] Sahoo HS, Silverman GM, Ingraham NE, Lupei MI, Puskarich MA, Finzel RL, Sartori J, Zhang R, Knoll BC, Liu S, Liu H, Melton GB, Tignanelli CJ, Pakhomov SVS. A fast, resource efficient, and reliable rule-based system for COVID-19 symptom identification. JAMIA Open. 2021 Aug;4(3):ooab070, doi:10.1093/jamiaopen/ooab070.

[9] Abdelwahab N, Ingraham NE, Nguyen N, Siegel L, Silverman G, Sahoo HS, Pakhomov S, Morse LR, Billings J, Usher MG, Melnik TE, Tignanelli CJ, Ikramuddin F. Predictors of Postacute Sequelae of COVID-19 Development and Rehabilitation: A Retrospective Study. Arch Phys Med Rehabil. 2022 Oct;103(10):2001-8, doi:10.1016/j.apmr.2022.04.009.

[10] CDC, Symptoms of Coronavirus, Centers for Disease Control and Prevention. 2020. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html (accessed Dec. 12, 2020).

[11] Pasc_acute_covid_lexicon. 2022. https://github.com/nlpie/lexical_gazetteer/tree/main/lexica/covid_pasc (accessed October 20, 2022).

[12] Brooke J. SUS-A quick and dirty usability scale. Usability Eval Ind. 1996 Jun;189(194):4-7.

[13] Pakhomov SV, Jacobsen SJ, Chute CG, Roger VL. Agreement between patient-reported symptoms and their documentation in the medical record. Am J Manag Care. 2008 Aug;14(8):530-9.

[14] Wang L, Foer D, MacPhaul E, Lo YC, Bates DW, Zhou L. PASCLex: A comprehensive post-acute sequelae of COVID-19 (PASC) symptom lexicon derived from electronic health record clinical notes. J Biomed Inform. 2022 Jan;125:103951, doi:10.1016/j.jbi.2021.103951.