Editorial: The National COVID Cohort Collaborative Consortium Combines Population Data with Machine Learning to Evaluate and Predict Risk Factors for the Severity of COVID-19

Dinah V. Parums, MD PhD

Science Editor, Medical Science Monitor, International Scientific Information, Inc., Melville, NY, USA
e-mail: dinah.v.parums@isi-science.com

Abstract   Infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that causes coronavirus disease 2019 (COVID-19) commonly presents with pneumonia. However, COVID-19 is now recognized to involve multiple organ systems with varying severity and duration. In July 2021, the findings from a retrospective population study from the National COVID Cohort Collaborative (N3C) Consortium were published that included analysis by machine learning methods of 174,568 adults with SARS-CoV-2 infection from 34 medical centers in the US. The study stratified patients for COVID-19 according to the World Health Organization (WHO) Clinical Progression Scale (CPS). Severe clinical outcomes were identified as the requirement for invasive ventilatory support, or extracorporeal membrane oxygenation (ECMO), and patient mortality. Machine learning analysis showed that the factor most strongly associated with severity of clinical course in patients with COVID-19 was pH. A separate multivariable logistic regression model showed that independent factors associated with more severe clinical outcomes included age, dementia, male gender, liver disease, and obesity. This Editorial aims to present the rationale and findings of the largest population cohort of adult patients with COVID-19 to date and highlights the importance of using large population studies with sophisticated analytical methods, including machine learning.

Keywords: **Editorial • Population Characteristics • Machine Learning • Epidemiology • Severe Acute Respiratory Syndrome • COVID-19 • Severe Acute Respiratory Syndrome Coronavirus 2**

Infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that causes coronavirus disease 2019 (COVID-19) commonly presents with pneumonia. However, COVID-19 is now recognized to involve multiple organ systems with varying severity and duration [1]. Early in the COVID-19 pandemic, attempts were made to classify the severity of COVID-19 based on symptom severity, degree of impaired lung function, and lung imaging changes on lung computed tomography (CT) [2]. However, impaired pulmonary function and imaging changes, such as consolidation and ground-glass opacities, are non-specific and associated with other types of viral pneumonia and infiltrative lung diseases [2,3]. The radiological term, ground-glass opacity, has been used for more than two decades to describe a mild increase in lung opacity [4]. However, the lung vasculature and bronchial structures remain visible, and this is not the same condition as consolidation [4].

The US National Institutes of Health (NIH) currently provides guidance to identify patients with severe COVID-19 pneumonia [5]. Severe COVID-19 includes the following clinical criteria: an oxygen saturation (SpO$_2$) of <94% on room air; a respiratory rate of >30 breaths per minute; a ratio of arterial oxygen partial pressure to fractional inspired oxygen (PaO$_2$/FiO$_2$) of <300 mmHg; or lung infiltrates that involve >50% of the lung on CT imaging [5].

In August 2020, the World Health Organization (WHO) developed the Clinical Progression Scale (CPS) to stratify patients for COVID-19 research studies according to a minimum set of three outcome measures [6]. The three WHO CPS elements include: a measure of SARS-CoV-2 viral burden determined by quantitative polymerase chain reaction (PCR); patient survival or mortality at hospital discharge or 60 days following initial diagnosis; and a measure of progression through the healthcare system using the WHO CPS criteria [6].

Standardized classification systems for disease severity are the basis for clinical studies, including clinical trials, that provide evidence for therapies, vaccines, and clinical guidelines. Studies from single centers or regions have evaluated clinical and laboratory findings in patients with confirmed COVID-19 to identify risk factors for disease severity and mortality in COVID-19 [7].

Patient age, immune suppression, and comorbidities, particularly chronic lung disease, were initially identified as risk factors for disease severity and patient mortality [7]. However, data from large population studies have been awaited to evaluate the clinical effects of SARS-CoV-2 infection and identify the risk factors for disease severity and patient mortality [8]. Because of the complex nature of COVID-19, large population studies will require complex analysis using machine learning to allow for the analysis of multiple confounders and risk factors over time [9].

In July 2021, the findings from a retrospective cohort study from the National COVID Cohort Collaborative (N3C) Consortium were published that included analysis by machine learning of 174,568 adults with SARS-CoV-2 infection from 34 medical centers in the US [10]. The National COVID Cohort Collaborative (N3C) contains members of the National Institutes of Health Clinical and Translational Science Awards Program, the IDeA Centers for Translational Research, TriNetX, the National Patient-Centered Clinical Research Network, the Observational Health Data Sciences and Informatics network, and the Accrual to Clinical Trials network [11]. The National COVID Cohort Collaborative (N3C) is a centralized, electronic health record repository representing the largest population cohort of adult patients with COVID-19 to date and includes 1,926,526 individuals [10,11].

In the National COVID Cohort Collaborative (N3C) Consortium study, patients were included between January 1, 2020, and December 7, 2020 [10]. More than 99% of patients in this database had a confirmed diagnosis of SARS-CoV-2 infection using polymerase chain reaction, with only <1% diagnosed by antigen testing alone [10]. A control population included SARS-CoV-2-negative individuals [10]. Patients were stratified using the WHO COVID-19 severity scale and patient demographic characteristics [6,10]. The National COVID Cohort Collaborative (N3C) used 64 data inputs taken on the first hospital day [10]. Machine learning models accurately predicted ultimate clinical severity using commonly collected clinical data from the first 24 hours of hospital admission [10]. Over time, data on the differences between the groups underwent multivariable logistic

regression analysis [10]. Random forest models and eXtreme Gradient Boosting (XGBoost) machine learning models were used to predict mortality and severe clinical outcomes that included the requirement for invasive ventilatory support or extracorporeal membrane oxygenation (ECMO) [10].

The findings from the National COVID Cohort Collaborative (N3C) Consortium of 174,568 SARS-CoV-2-positive adults had a mean age of 44.4 ±18.6 years, of which 53.2% were women, 32,472 (18.6%) were hospitalized, and of the patients hospitalized, 6,565 (20.2%) patients had a severe clinical course [10]. The overall mortality rate was 11.6% [10]. Patient overall mortality decreased from 16.4% in March and April 2020 to 8.6% in September and October 2020 [10].

Machine learning analysis showed that the factor most strongly associated with severity of clinical course in patients with COVID-19 was pH, which was consistent between machine learning methods [10]. The results from a separate multivariable logistic regression model showed that the following were independent factors associated with a more severe clinical outcomes: age (OR, 1.03; 95% CI, 1.03-1.04); dementia (OR, 1.26; 95% CI, 1.13-1.41); male gender (OR, 1.60; 95% CI, 1.51-1.69); liver disease (OR, 1.20; 95% CI, 1.08-1.34); and obesity (OR, 1.36; 95% CI, 1.27-1.46) [10]. The study also showed racial disparity for disease severity, with African American and Asian American patients having more severe clinical outcomes [10].

Conclusions

Machine learning analysis of a national multicenter population database from the US the National COVID Cohort Collaborative (N3C) Consortium analyzed clinical and demographic data on 174,568 SARS-CoV-2-positive adults admitted to hospital. This study confirmed that patient mortality from COVID-19 decreased during 2020. Patient demographic characteristics of age, male gender, and comorbidities, including obesity, were associated with increased clinical severity.

References:

1. Parums DV. Editorial: Long COVID, or post-COVID syndrome, and the global impact on health care. Med Sci Monit 2021;27:e933446

2. Li L, Sun W, Han M, et al. A study on the predictors of disease severity of COVID-19. Med Sci Monit. 2020;26:e927167

3. Hochhegger B, Zanon M, Altmayer S, et al. COVID-19 mimics on chest CT: A pictorial review and radiologic guide. Br J Radiol. 2021;94(1118):20200703

4. Collins J, Stern EJ. Ground-glass opacity at CT: The ABCs. Am J Roentgenol 1997;169:355-67

5. National Institutes of Health (NIH). COVID-19 Treatment Guidelines Panel. Clinical spectrum of SARS-CoV-2 infection. April 21, 2021. Available at: https://www.covid19treatmentguidelines.nih.gov/overview/clinical-spectrum/

6. WHO Working Group on the Clinical Characterization and Management of COVID-19 infection. A minimal common outcome measure set for COVID-19 clinical research. Lancet Infect Dis. 2020;20(8):e192-97

7. Wang Y, Yao S, Liu X, et al. Risk factors of coronavirus disease 2019-related mortality and optimal treatment regimens: A retrospective study. Med Sci Monit. 2021;27:e926751

8. Parums DV. Editorial: Registries and population databases in clinical research and practice. Med Sci Monit. 2021;27:e933554

9. Parums DV. Editorial: Artificial intelligence (AI) in clinical medicine and the 2020 CONSORT-AI study guidelines. Med Sci Monit. 2021;27:e933675

10. Bennett TD, Moffitt RA, Hajagos JG, et al.; National COVID Cohort Collaborative (N3C) Consortium. Clinical characterization and prediction of clinical severity of SARS-CoV-2 infection among US adults using data from the US National COVID Cohort Collaborative. JAMA Netw Open. 2021;4(7):e2116901

11. Haendel MA, Chute CG, Bennett TD, et al; N3C Consortium. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021;28(3):427-43