Research team creates statistical model to predict COVID-19 resistance

February 22 2023

A research team from Johns Hopkins Medicine and The Johns Hopkins University has created and preliminarily tested a machine-learning statistical model that—using data from electronic health records—may soon be able to predict who is naturally resistant to infection by SARS-CoV-2, the virus that causes COVID-19. Credit: M.E. Newman, Johns Hopkins Medicine, using public domain images, including electron micrograph of SARS-CoV-2 infection courtesy of the National Institute of Allergy and Infectious Diseases

Researchers from Johns Hopkins Medicine and The Johns Hopkins University have created and preliminarily tested what they believe may be one of the first models for predicting who has the highest probability of being resistant to COVID-19 in spite of exposure to SARS-CoV-2, the virus that causes it.

The study is reported online today in the journal *PLOS ONE.*

"If we can identify which people are naturally able to avoid infection by SARS-CoV-2, we may be able to learn—in addition to societal and behavioral factors—which genetic and environmental differences influence their defense against the virus," says lead study author Karen (Kai-Wen) Yang, a biomedical engineering graduate student in the Translational Informatics Research and Innovation Lab at The Johns Hopkins University. "That insight could lead to new preventive measures and more highly targeted treatments."

For its study, the research team set out to determine whether a machine-learning statistical model could use health characteristics stored in electronic health records—providing patient data such as comorbidities (other medical conditions) and prescribed medications—as a means to pinpoint people with a natural ability to avoid SARS-CoV-2 infection. Those persons, says Yang, could then be studied to better understand the factors enabling their resistance.

A machine-learning model is a computer program or system that uses mathematical algorithms to find statistical patterns, and then apply the patterns moving forward. This gives such systems the ability to imitate human thinking and reasoning, and similar to the brain, learn over time.

"Using a machine-learning system to recognize complex patterns in large numbers of people with COVID-19 enabled another team of Johns Hopkins Medicine researchers in 2021 to predict the course of an

individual patient's case and determine the likelihood that it would become severe," says co-senior study author Stuart Ray, M.D., vice chair of medicine for data integrity and analytics, and professor of medicine at the Johns Hopkins University School of Medicine. "Based on their success, our team wondered if the same approach also might be applied to predicting who could be exposed to SARS-CoV-2 in close quarters and still not get infected."

To demonstrate the model's ability to predict COVID-19 resistance, the researchers first acquired data from a clinical registry called the Johns Hopkins COVID-19 Precision Medicine Analytics Platform Registry (JH-CROWN). The registry contains information for patients seen within the Johns Hopkins Health System who have been suspected of or confirmed as having a SARS-CoV-2 infection.

For their resistance study, the researchers only included individuals who received a COVID-19 test between June 10, 2020, and Dec. 15, 2020, and who reported "potential exposure to the virus" as the reason for testing.

The ending date was the point at which large-scale COVID-19 vaccination efforts started in the United States. Choosing this date, the researchers say, enabled them to avoid the effects on their findings of vaccines preventing infection rather than natural resistance.

The 8,536 study participants who reported exposure as their reason for getting COVID tested were divided into two groups: those who did not share a residence (called a "household" in this study) with any COVID-19 patients or whose residence had 10 or more patients; and those who shared a residence with 10 or fewer people, with at least one being a COVID-19 patient. The first group, with 8,476 of the participants, was designated as the Training and Testing Set, while the second group, called the Household Index (HHI) Set, had 60 members,

and was used as a separate testing set.

Keeping the household number to 10 or fewer, the researchers say, excluded people living in apartment complexes, dormitories and other higher-density, multi-unit living areas where exposure to a particular person positive for SARS-CoV-2 would be less intense.

To identify patterns and cluster participants so that those naturally resistant to SARS-CoV-2 stand out, both study sets were analyzed using the Maximal-frequent All-confident pattern Selection Pattern-based Clustering (MASPC) algorithm. MASPC is specifically designed for electronic health record data analysis that combines patient demographic information (age, sex and race), the International Statistical Classification of Diseases and Related Health Problems (ICD) medical diagnostic codes relevant to each case, outpatient medication orders and the number of comorbidities (other diseases) present.

"We hypothesized that MASPC would enable us to cluster patients with similar patterns in their data to define them as resistant and non-resistant to SARS-CoV-2, and with the hope that the algorithm would learn with each analysis how to improve the accuracy and reliability of future assignments," says Ray. "This initial study using JH-CROWN data was conducted to give life to that hypothesis, a proof-of-concept trial of our statistical model to show that resistance to COVID-19 might be predictable based a patient's clinical and demographic profile."

"In the Training and Testing Set, we identified 56 patterns of ICD codes split into two groups: associated with resistance or not associated," Yang says. "Statistical analyses of how well these patterns differentiated between resistance and non-resistance yielded five patterns that did the best job for our small and localized [Baltimore-Washington, D.C., metroplex] study population to define who was most likely exposed to SARS-CoV-2."

"Looking for these patterns in HHI Set—the individuals most likely to have been exposed to SARS-CoV-2 in close quarters—and then statistically analyzing the results, our model's best performance was 0.61," says Ray. "Since a score of 0.5 shows only chance association between the prediction and reality, and 1 is 100% association, this shows the model has promise as a tool for identifying people with COVID-19 resistance who can be further studied," says Ray.

Limitations to the study, says Ray, include potential bias from self-reporting of COVID-19 exposure by participants, the small number of participants in the HHI group, the possibility that participants tested for SARS-CoV-2 using home kits or at facilities outside the Johns Hopkins system (and therefore, the tests were not recorded in the JH-CROWN database), and the short timeframe of the study itself. He adds that future trails using national patient data are needed to validate the model's ability.

Along with Yang and Ray, the members of the study team from Johns Hopkins Medicine and Johns Hopkins University are graduate and undergraduate students Yijia Chen, Jacob Desman, Kevin Gorman, Chloé Paris, Ilia Rattsev, Tony Wei and Rebecca Yoo; and faculty co-senior authors Joseph Greenstein and Casey Overby Taylor.

More information: Kai-Wen K. Yang et al, Factors associated with resistance to SARS-CoV-2 infection discovered using large-scale medical record data and machine learning, *PLOS ONE* (2023). DOI: 10.1371/journal.pone.0278466

Provided by Johns Hopkins University School of Medicine

Citation: Research team creates statistical model to predict COVID-19 resistance (2023,

February 22) retrieved 4 April 2023 from https://medicalxpress.com/news/2023-02-team-statistical-covid-resistance.html