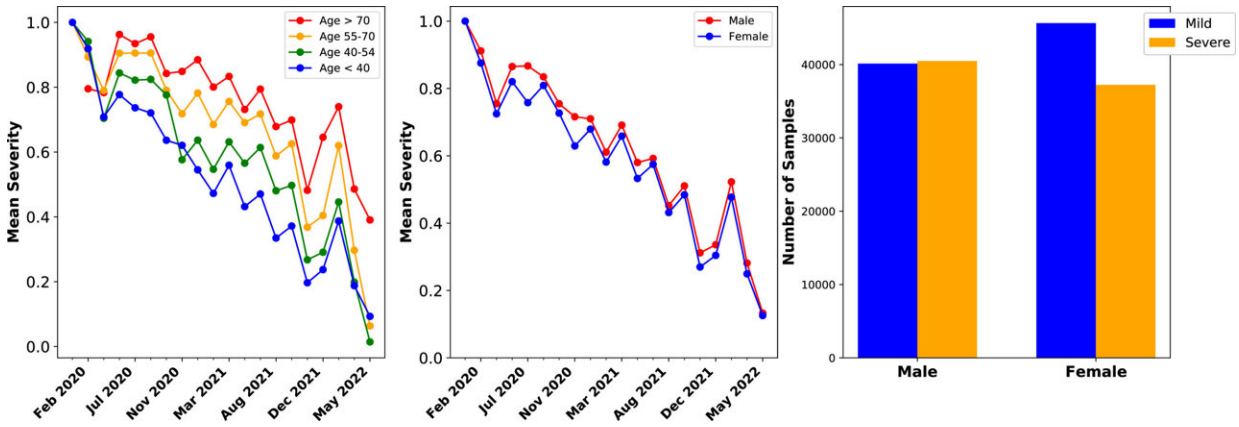
COVID radar: Genetic sequencing can help predict severity of next variant

September 1 2022



Patient age and gender metadata trends in GISAID data. (A – Left) Mean clinical severity over time for patients in different age groups, showing that the overall trends are generally consistent across age groups, with older patients having mean severity as shown in Panel A. (B – Middle) Mean clinical severity, separating male and female samples, showing consistent trends across gender with male patients generally having a somewhat higher ratio of severe cases. (C - Right) Number of mild and severe cases across all samples split by gender, showing that there are more mild cases than severe among samples from female patients. Credit: *Computers in Biology and Medicine* (2022). DOI: 10.1016/j.compbiomed.2022.105969

As public health officials around the world contend with the latest surge of the COVID-19 pandemic, researchers at Drexel University have created a computer model that could help them be better prepared for

the next one. Using machine learning algorithms, trained to identify correlations between changes in the genetic sequence of the COVID-19 virus and upticks in transmission, hospitalizations and deaths, the model can provide an early warning about the severity of new variants.

More than two years into the pandemic, scientists and public health officials are doing their best to predict how mutations of the SARS-CoV-2 virus are likely to make it more transmissible, evasive to the immune system and likely to cause severe infections. But collecting and analyzing the genetic data to identify new variants—and linking it to the specific patients who have been sickened by it—is still an arduous process.

Because of this, most public health projections about new "variants of concern"—as the World Health Organization categorizes them—are based on surveillance testing and observation of the regions where they are already spreading.

"The speed with which new variants, like omicron have made their way around the globe means that by the time public health officials have a good handle on how vulnerable their population might be, the virus has already arrived," said Bahrad A. Sokhansanj, Ph.D., an assistant research professor in Drexel's College of Engineering who led development of the computer model. "We're trying to give them an early warning system—like advanced weather modeling for meteorologists—so they can quickly predict how dangerous a new variant is likely to be—and prepare accordingly."

The Drexel model, which was recently published in the journal *Computers in Biology and Medicine*, is driven by a targeted analysis of the genetic sequence of the virus's spike protein—the part of the virus that allows it to evade the immune system and infect healthy cells, it is also the part known to have mutated most frequently throughout the

pandemic—combined with a mixed effects machine learning analysis of factors such as age, sex and geographic location of COVID patients.

Learning to find patterns

The research team used a newly developed machine learning algorithm, called GPBoost, based on methods commonly used by large companies to analyze sales data. Via a textual analysis, the program can quickly home in on the areas of the genetic sequence that are most likely to be linked to changes in the severity of the variant.

It layers these patterns with those that it gleans from a separate perusal of patient metadata (age and sex) and medical outcomes (mild cases, hospitalizations, deaths). The algorithm also accounts for, and attempts to remove, biases due to how different countries collect data. This training process not only allows the program to validate the predictions it has already made about existing variant, but it also prepares the model to make projections when it comes across new mutations in the spike protein. It shows these projections as a range of severity—from mild cases to hospitalizations and deaths—depending on the age, or sex of a patient.

"When we get a sequence, we can make a prediction about risk of severe disease from a variant before labs run experiments with animal models or cell culture, or before enough people get sick that you can collect epidemiological data. In other words, our model is more like an early warning system for emerging variants," Sokhansanj said.

Genetic and patient data from the GISAID database—the largest compendium of information on people who have been infected with the coronavirus—were used to train the algorithm. Once the algorithms were primed the team used them to make projections about the omicron subvariants post-BA.1 and BA.2.

"We show that future omicron subvariants are likelier to cause more severe disease," Sokhansanj said. "Of course, in the real world, that increased disease severity will be mitigated by prior infection by the previous omicron variants—this factor is also reflected in the modeling."

Keeping up with COVID

Drexel's targeted approach to predictive modeling of COVID-19 is a crucial development because the massive amount of genetic sequencing data being collected has strained standard analysis methods to extract useful information quickly enough to keep up with the virus's new mutations.

"The amount of spike protein mutations has already been quite substantial and it will likely continue because the virus is encountering hosts that have never been infected before," said Gail Rosen, Ph.D., a professor in the College of Engineering, who heads Drexel's Ecological and Evolutionary Signal-processing and Informatics Laboratory.

"Some estimates suggest that SARS-CoV-2 has only 'explored' as little as 30-40% of the potential space for spike mutations," she said. "When you consider that each mutation could impact key virus properties, like virulence and immune evasion, it seems vital to be able to quickly identify these variations and understand what they mean for those who are vulnerable to infection."

Rosen's lab has been at the forefront of using algorithms to cut though the noise of genetic sequencing data and identify patterns that are likely to be significant. Early in the pandemic the group was able to track the geographic evolution of new SARS-CoV-2 variants by developing a method for quickly identify and labeling its mutations. Her team has continued to leverage this process to better understand the patterns of the pandemic.

Vision among variables

Up until now, scientists have predominantly used genetic sequencing to better identify mutations alongside lab experiments and epidemiological studies. There has been little success in linking specific genetic sequence variations to virality of new variants. The Drexel researchers believe this is due to progressive changes in vaccination and immunity over time, as well as variations in how data is reported in different countries.

"We know that each successive COVID-19 variant thus far has resulted in slightly milder infections because of increases in vaccination, immunity and health care providers having a better understanding of how to treat infections. But what we have discovered through our mixed effects analysis is that this trend does not necessarily hold for each country. This is why our model considers geographic location as one of the variables taken into consideration by the machine learning algorithm," Sokhansanj said.

While disparities and inconsistencies in patient and public health data have been a challenge for public health officials throughout the pandemic, the Drexel model is able to account for this and explain how it affected the algorithm's projections.

"One of our key goals was making sure that the model is explainable, that is, we can tell why it's making the predictions that it's making," Sokhansanj said. "You really want a model that allows you to look under the hood to see, for example, the reasons why its predictions may or may not agree with what biologists understand from lab experiments—to ensure the predictions are built on the right structure."

A better view

The team notes that advances like this underscore the need to provide

more public health resources to vulnerable areas of the world—not only for treatment and vaccination, but also for collecting public health data, including sequencing emerging variants.

The researchers are currently using the model to more rigorously analyze the current group of emerging variants that will become dominant after omicron BA.4 and BA.5.

"The virus can and will continue to surprise us," Sokhansanj said. "We urgently need to expand our global capacity to sequence variants, so that we can analyze the sequences of potentially dangerous variants as soon as they show up—before they become a worldwide problem."

More information: Bahrad A. Sokhansanj et al, Predicting COVID-19 disease severity from SARS-CoV-2 spike protein sequence by mixed effects machine learning, *Computers in Biology and Medicine* (2022). DOI: 10.1016/j.compbiomed.2022.105969

Provided by Drexel University