How a U.S. COVID-19 Data Registry Fuels Global Research

Farhana Nakhooda	SVP, Health Catalyst, Asia Pacific (APAC)	Health Catalyst
Praveen Deorani	Sr. Data Scientist, Singapore Ministry of Healthcare	Health Catalyst
Larry Lofgreen	Asia Pacific Sales and Solutions Consulting, VP	Health Catalyst
Sadiqa Mahmood, DDS, MPH	General Manager & SVP, Life Sciences Business	Health Catalyst
Pol Margalef, PhD	Strategy & Business Dev Consultant, Life Sciences	Health Catalyst

The COVID-19 outbreak has been a significant U.S. and global concern, given the speed of spread and breadth of health impacts (both known and unknown) on the population level. The virus causes fever, cough, lack of smell, fatigue, and mild to severe respiratory complications, which, if very severe, can lead to patient death. Meanwhile, incomplete, non-transparent, and out-of-date COVID-19 data is one of the main barriers to understanding and managing the virus nationally and abroad, as well as developing a vaccine. To circumvent the lack of real-world, research-grade evidence, researchers are looking to innovative sources of comprehensive, real-time COVID-19 data.

A national COVID-19 data set that leverages deep aggregated EMR data delivers the depth and breadth of understanding researchers need to manage the virus and develop a vaccine. The Health Catalyst Touchstone® COVID-19 Registry and Insights, for example, includes de-identified data from 80 million patients across the United States and tracking data from three national sources—Johns Hopkins University, the New York Times, and The COVID Tracking Project. With such broad data access, data analysts can leverage data on a national scale to drive population-level insights about surveillance, testing, capacity planning, and treatment response.

National-Level COVID-19 Data Powers Global Research

Touchstone and the national COVID-19 registry also promise to inform research beyond U.S. borders. In the summer of 2020, the Singapore Ministry of Healthcare's (MOH)

Office for Healthcare Transformation (MOHT), in collaboration with Health Catalyst, used Touchstone COVID-19 data to develop a machine learning tool that helps predict the likelihood of COVID-19 mortality—a critical insight for driving care to highest-risk patients and managing the outbreak on a population level. To validate the accuracy of their predictive tool, Health Catalyst compared its results with results published in the literature and determined its registry-informed research aligned closely to peer-reviewed publications.

“For a rapidly evolving situation like COVID-19, medical researchers can’t rely solely on clinical trials for guidance,” explains Praveen Deorani, Senior Data Scientist, for the Singapore MOHT. “As a practical alternative to informing medical decisions, a machine learning model can generate and analyze real-world evidence much faster.”

Registry-Driven Analytics Tools Leverage COVID-19 Data for Decision Support

In an effort to assist neighboring countries that may not have the research resources available, the Singapore MOHT sought to provide analytic tools to assist in managing the pandemic. However, Singapore’s population size and the strict control measures implemented in Singapore combined to limit both the nation’s number of COVID-19 cases and the COVID-19 mortality rate, leaving a dearth of data to power predictive tools.

Data scientists with the Singapore MOHT evaluated detailed COVID-19 data from the Touchstone registry to identify patient factors linked to COVID-19 mortality (Figure 1).

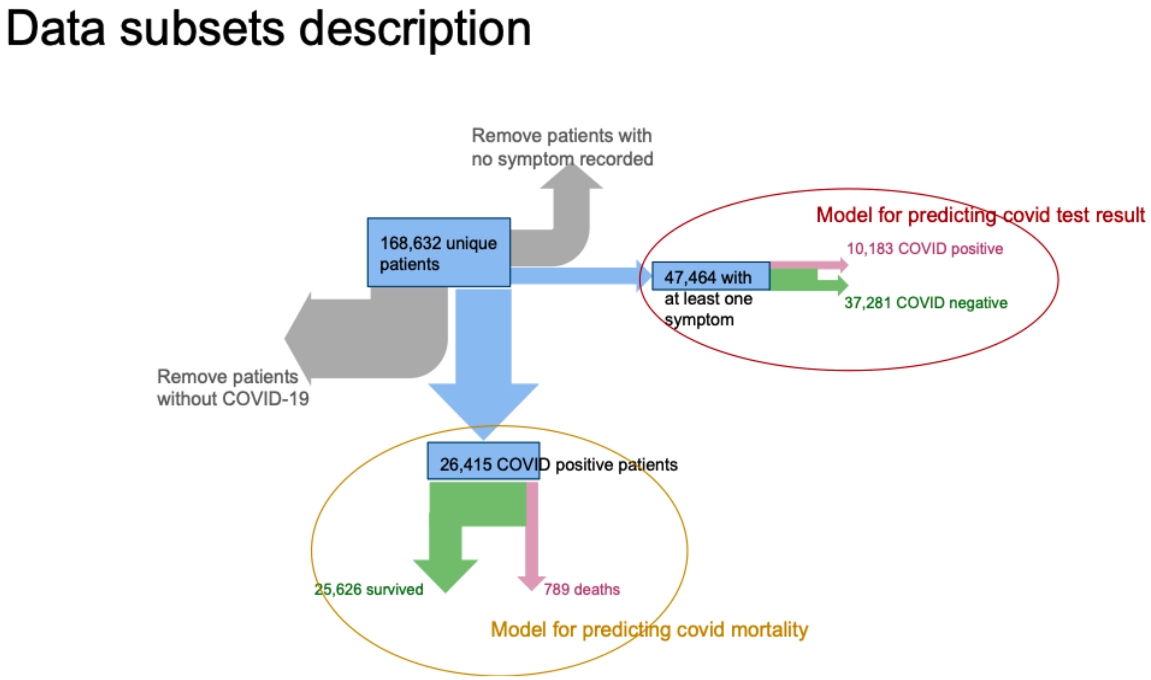
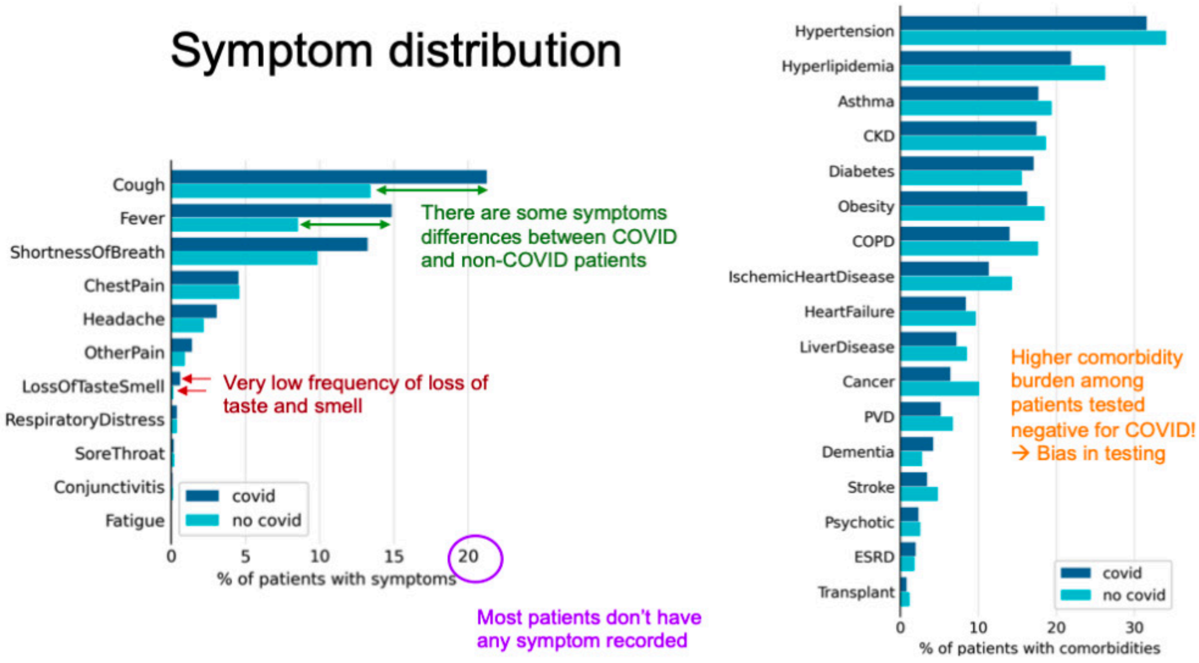

Data subsets description
Remove patients with no symptom recorded
168,632 unique patients
Model for predicting covid test result
10,183 COVID positive
47,464 with at least one symptom
37,281 COVID negative
Remove patients without COVID-19
26,415 COVID positive patients
25,626 survived
789 deaths
Model for predicting covid mortality

Figure 1: Factors linked to COVID-19 mortality.

The Touchstone COVID-19 data set contained deidentified data for 168,632 unique patients. For comparison purposes, this dataset included patients with COVID-19-related symptoms and diagnoses. Of these unique patients, 47,464 exhibited at least one COVID-19-related symptom, approximately 21 percent of whom tested positive for COVID-19. Similarly, the data contained 26,415 patients who tested positive for COVID-19 (61 percent were either asymptomatic, or the treating facility didn't document the symptoms). The COVID-19-related mortality rate for COVID-19-positive patients was approximately 3 percent (789 of 26,415).

The initial analysis effort focused on providing a triage tool for prioritizing care of patients exhibiting COVID-19-related symptoms. As Figure 2 shows, patients who tested positive for COVID-19 had different symptom distributions versus those who did not test positive. However, most patients were either asymptomatic or had no symptoms recorded. The small number of patients exhibiting loss of taste/smell is of particular interest to the MOHT, as this symptom has been seen as a strong indicator of COVID-19 in Singapore.


Symptom distribution

There are some symptoms differences between COVID and non-COVID patients

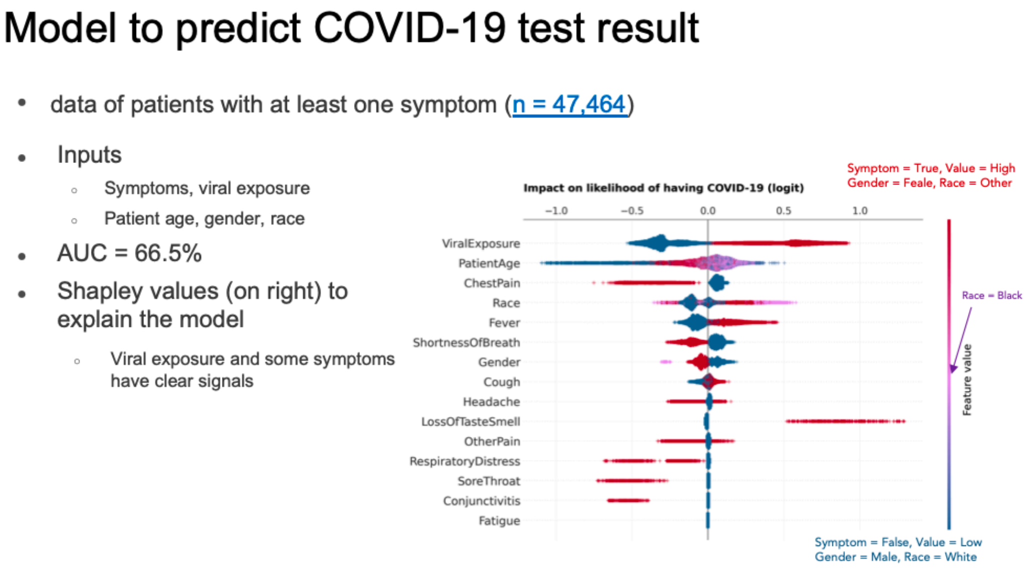
Very low frequency of loss of taste and smell

Most patients don't have any symptom recorded

Higher comorbidity burden among patients tested negative for COVID! → Bias in testing

Figure 2: Symptom distribution for patients with COVID-19.

Despite the general lack of symptom data, when the MOHT researchers compared the correlation of symptoms to a positive COVID-19 test, two symptoms stood out: prior viral exposure and loss of taste/smell (the latter confirming what Singapore had determined through their testing regimes). Ultimately, the U.S. symptom data was too sparse to form the basis of a predictive model that could perform better than the literature-based, deterministic test result model that MOHT had already developed (Figure 3).



Model to predict COVID-19 test result

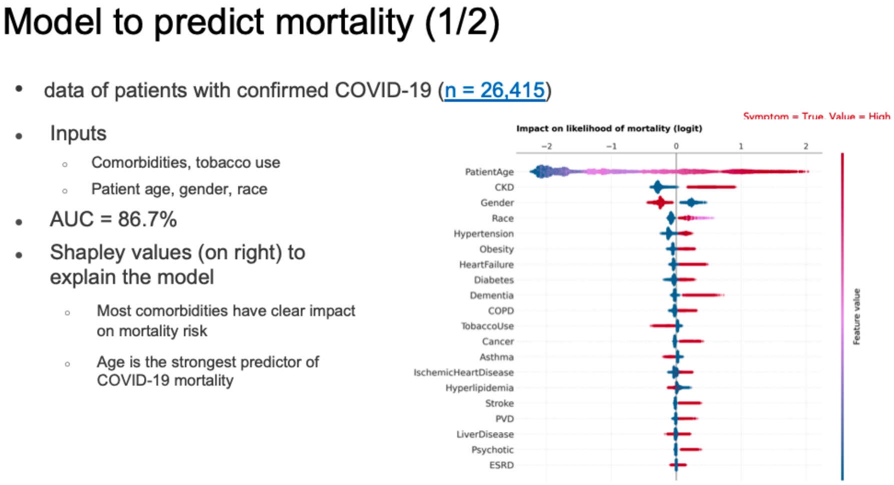
- data of patients with at least one symptom (n = 47,464)
- Inputs
  - Symptoms, viral exposure
  - Patient age, gender, race
- AUC = 66.5%
- Shapley values (on right) to explain the model
  - Viral exposure and some symptoms have clear signals

Figure 3: The MOHT COVID-19 test result prediection model.

A Data-Informed Machine Learning Tool Helps Predict Who Is Likely to Die of COVID-19

After the MOHT initial analysis efforts, the organization used factors such as age, race, gender, and comorbidities (including hypertension, cancer, and more), to produce a machine learning prediction tool to help clinicians identify COVID-19 patients at the highest risk of death (Figure 4). Some of the MOHT's most meaningful insights include the following:

- The mortality rate varies significantly by age group and gender and somewhat with race.
- Black or African Americans have higher mortality rates despite a slightly lower age.
- The mortality rate depends on comorbidities independent of age distribution.



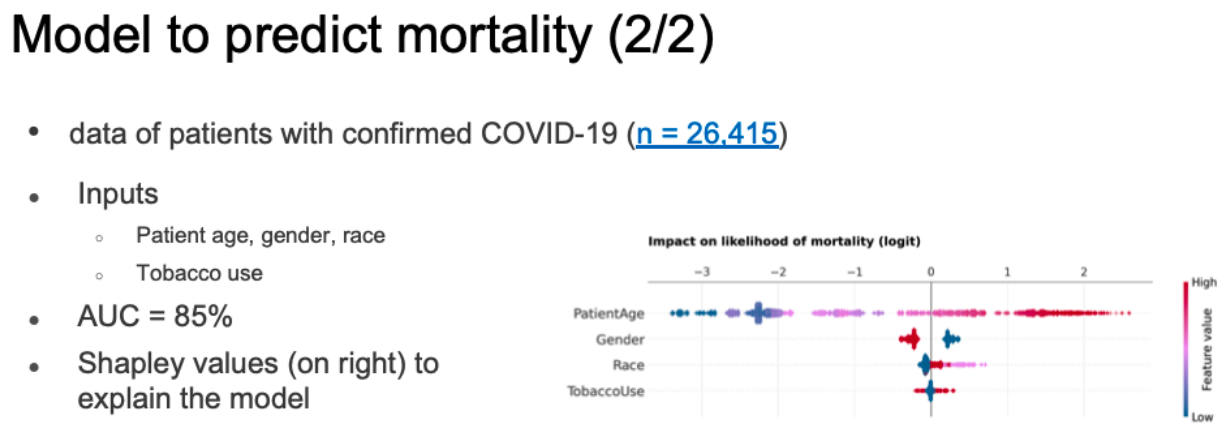
Model to predict mortality (1/2)

- data of patients with confirmed COVID-19 (n = 26,415)
- Inputs
  - Comorbidities, tobacco use
  - Patient age, gender, race
- AUC = 86.7%
- Shapley values (on right) to explain the model
  - Most comorbidities have clear impact on mortality risk
  - Age is the strongest predictor of COVID-19 mortality

Figure 4: The MOHT COVID-19 mortality prediction tool 1/2.

In contrast to the lack of symptom data captured, patient demographic and comorbidity data supported a mortality prediction model (an aggregate measure of performance across all possible classification thresholds, an AUC, of 86.7 percent). For the comorbidities in the chart above, red indicates existence of the condition, and blue indicates absence of the condition. As the values show, most comorbidities have an obvious impact on mortality risk.

However, comorbidity-based prediction is only useful if the analysts know a patient's comorbidities. Therefore, given the observed impact of age, gender, and race in the comorbidity-based model, the MOHT data scientists created a second model using only those features likely universally available to clinicians: age, gender, race, and history of tobacco use. As Figure 5 shows, this model was performed nearly the same as the model with comorbidities (an AUC 85 percent versus the original AUC of 86.7 percent).



Model to predict mortality (2/2)

- data of patients with confirmed COVID-19 (n = 26,415)
- Inputs
  - Patient age, gender, race
  - Tobacco use
- AUC = 85%
- Shapley values (on right) to explain the model

Figure 5: MOHT COVID-19 mortality prediction tool 2/2.

The COVID-19 Mortality Prediction Model Stands up to Peer-Reviewed Literature

To verify the accuracy of the COVID-19 mortality prediction model, the MOHT reviewed published literature to compare the model's outcomes with other research. The team determined its prediction model results were overwhelmingly consistent with other peer-reviewed studies.

The following lists offer examples of factors the MOHT model uses to predict COVID-19 mortality and some of the published literature that confirms their relationship to COVID-19 mortality:

- Patient age—Several studies indicate patient age is a reliable predictor for COVID-19 mortality:
  - According to a study in China, patients aged 65 years and older have a higher risk of COVID-19 mortality.
  - Researchers in Korea find a higher rate of COVID-19 death in older adults.
- Race—Race has some mixed results as a predictor of COVID-19 death, but some studies show a correlation:

    - New York study determines an association between race with COVID-19 mortality.
    - The U.S. Black population has a higher rate of COVID-19 case fatality.
- Gender—Studies associate gender with COVID-19 mortality:
    - In China, men are more at risk for having the worst COVID-19 outcomes and death.
    - A multivariate regression identifies being male as a risk factor for COVID-19 mortality.
- Cancer—Patients with COVID-19 and cancer have a greater risk of death.
    - Age, gender, and comorbidities drive the risk of COVId-19 death among patients with cancer.
    - Due to unfavorable prognostic factors, hospitalized patients with cancer and COVID-19 had a high case-fatality rate.

Partnering for Meaningful COVID-19 Understanding

One of the most promising uses of these COVID-19-data-drive prediction models may be in prioritization of viral testing in localities with insufficient resources. The first priority would be the allocation of COVID-19 tests to frontline healthcare workers and individuals in contact with a large number of people, such as cashiers and bus drivers. For the remaining population, the thresholds of risk for COVID-19 (given symptoms) and risk of death from the virus could determine test allocation. Similarly, these data-powered models may support early allocation of vaccines when they becomes available, as immunization among high-risk individuals maximizes the early impact of a vaccine.

Combining the Touchstone COVID-19 Registry and Insights aggregated data from U.S. healthcare providers with the expertise and experience of Singapore’s MOHT provided capability and insights neither organization could muster alone. The opportunities for global collaborations such as this are endless and create a huge opportunity for the research community at large to leverage real-world evidence to address global health issues and ultimately improve health outcomes.