CNBC

Over 30 million people told Facebook if they had the coronavirus or wore masks — and now it will be used for science

By Christina Farr
FRI, SEP 4 2020 UPDATED FRI, SEP 4

When Carnegie Mellon University researchers had the idea to put together a survey asking the general public about their coronavirus symptoms, the scientists knew they needed to collect millions of data points to learn anything meaningful.

So they asked Facebook, which has a public team that specializes in using analytics for humanitarian causes called "Data for Good," for its help.

The survey, which went live to Facebook's billions of users about six months ago, has so far collected data from more than 30 million people around the world. The survey asks whether they tested positive for the virus, if they wear masks and practice socially distancing as well as if they're currently experiencing symptoms. Respondents also share data about their demographics, like their age, as well as their mental health status and preexisting medical conditions.

More than 1.5 million people fill out the survey each week. To preserve privacy, Facebook said it doesn't have direct access to the responses. Carnegie Mellon has now published aggregated data through its COVIDcast API, as well as real-time visualizations.

But there's still a few big questions to be answered: Will this data be truly useful? And can it predict the next outbreak of Covid-19 before it happens?

To find out, a group of epidemiologists and infectious disease experts from Carnegie Mellon, the University of Maryland, the Duke Margolis Center for Health Policy and Resolve to Save Lives, a nonprofit headed up by former CDC director Tom Frieden, have launched a challenge that's open to any data scientist or researcher.

With prize money funded by Facebook, the ultimate goal is to see if the dataset can be used to help find the next Covid-19 surge, so public health officials can deploy scarce resources accordingly.

“It’s a wealth of information I’ve been stunned isn’t in broader use,” Dr. Farzad Mostashari, the former national coordinator for health information technology at the Department of Health and Human Services, said in a phone interview. He also helped create the challenge. “If it’s better understood, this could be a big step forward.”

Once submissions are received — the first deadline is Sept. 29 — a scientific committee of epidemiologists and data scientists will review them. Mostashari, Boston Children’s Hospital’s John Brownstein and Alex Reinhart, an assistant teaching professor in statistics and data science at Carnegie Mellon, are on the committee, along with about a dozen others working on the frontlines of the pandemic.

Google flu

The idea of using consumer technology tools like Facebook and Google to collect information about disease is nothing new.

In the mid-2000s, a group of epidemiologists, including Brownstein, started working with tech companies to figure out whether their data could be used to advance public health programs. That resulted in projects like Google Flu Trends, started in 2008, which aimed to use search trends to figure out the prevalence of influenza in specific regions.

Google Flu Trends wasn’t a huge success story in the end, in part because Google learned too late that these datasets needed to be combined with information collected by public health agencies like the CDC. It folded in the summer of 2015.

But researchers still see the benefit in collecting information on people’s symptoms, whether it comes from search terms, online surveys or wearable devices. Combined with other so-called “syndromic surveillance” datasets, such as how many patients are reporting influenza-like illness in emergency rooms, the data collected by tech companies can help predict epidemics, researchers say.

Symptom searches

Covid-19 has now inspired many of the biggest tech companies to once again get behind funding collaborations with public health departments, after learning from their past failures.

“We’ve validated this kind of data over time,” said Brownstein, who continues to work with tech giants including Facebook, Google and Uber. “And now, the tech companies are putting significant resources behind it.”

Also this week, Google shared that it is exploring whether symptom search trends, such as searches for fever, can predict a potential Covid-19 outbreak and help researchers map the spread of the virus. The approach is similar to Google Flu Trends, but the company said it is looking for feedback from public health researchers to make the dataset more useful over time.

In Mostashari’s view, there is a need for these new kinds of datasets because the current methodologies are far from perfect. Because of insufficient testing in countries like the United States, it’s a challenge for public health departments to glean accurate case counts. Deaths are, of course, a lagging indicator. And surveilling hospital emergency rooms alone can be insufficient, because of changes in how people seek care. For instance, in a pandemic, fewer people in general are going to emergency rooms than normal — and that can impact the data.

‘The cat’s pajamas’

Mostashari said the survey data might have helped researchers predict the recent surge of cases in Florida. “There’s enough evidence to suggest it could be a big deal,” he said.

Other researchers agree. “It was clear within say a few months of gathering the data that the signal seemed to have some correlation with confirmed case counts,” added Carnegie Mellon’s Reinhart, referring to his group’s initial efforts to see if the symptom data correlated with state-by-state reports on the number of cases. “It’s taken us longer though to do a deeper analysis given the sample size.”

But Reinhart and Mostashari say they are open to being proven wrong. They are hoping that the researchers who join the challenge will test their assumptions and unearth yet more insights along the way.

“We want it to be ripped apart,” said Mostashari. “And for those who submit to the challenge to ask questions about whether this (dataset) is truly the cat’s pajamas, or if we’re seeing correlation without causation.”