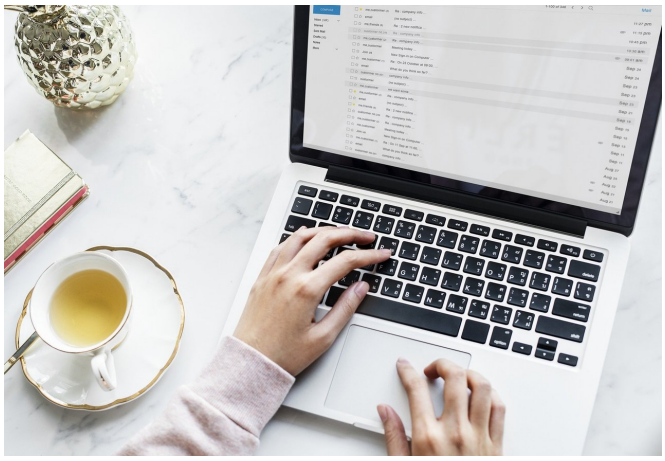
Researchers use search engines, social media to predict syphilis trends

16 April 2018, by Enrique Rivero



Credit: CC0 Public Domain

UCLA-led research finds that internet search terms and tweets related to sexual risk behaviors can predict when and where syphilis trends will occur.

Two studies from the UCLA-based University of California Institute for Prediction Technology, in collaboration with the Centers for Disease Control and Prevention, or CDC, found an association between certain risk-related terms that Google and Twitter users researched or tweeted about and subsequent syphilis trends that were reported to the CDC. The researchers were able to pinpoint these cases at state or county levels, depending on the platform used.

"Many of the most significant public health problems in our society today—HIV and sexually transmitted infections, opioid abuse and cancer—could be prevented if we had better data on when and where these issues were occurring," said Sean Young, founder and director of the UCLA Center for Digital Behavior and the UC Institute for Prediction Technology. "These two studies suggest that social media and internet search data might help to fix this problem by predicting when and where future syphilis cases may occur. This could be a tool that government agencies such as the CDC might use," added Young, who is also an associate professor of family medicine at the David Geffen School of Medicine at UCLA.

One study, to be published in the peer-reviewed journal *Epidemiology*, investigated the association between state-level search queries on Google with primary and secondary syphilis cases—the earliest and most transmissible stages in the sexually transmitted infection—that were subsequently reported in these states.

For this study, the researchers compiled data for 25 keywords and phrases (such as "find sex" and "STD") collected on Google Trends from Jan. 1, 2012, to Dec. 31, 2014. They also obtained weekly county-level syphilis data from the CDC covering the same time period for all 50 states, merged that data by state and collated them with the weekly Google Trends data they had collected.

The research incorporated a type of statistical computer science model called machine learning, which can look through large amounts of data to find patterns and predict those patterns. This artificial intelligence-based machine looked at the relationship between people's syphilis-related searches on Google and actual rates of syphilis over a period of time. After learning that pattern, it tested whether it could accurately predict future syphilis cases by using just the syphilis-related Google search terms.

Researchers found that the model predicted 144 weeks of syphilis counts for each state with 90 percent accuracy, allowing them to predict state-level trends in syphilis before they would have occurred.

Researchers from the institute found the same held

true with Twitter. In a study published in *Preventive Medicine*, they took county-level Twitter data from May 26 to Dec. 9, 2012, amounting to 8,538 geo-located tweets. As with the Google Trends analysis, the researchers compiled a list of words associated with sexual risk behaviors.

They reviewed weekly county-level cases of primary and secondary syphilis and early latent syphilis (infection within the previous 12 months, with no symptoms evident) that likely occurred over the previous 12 months. The cases were from the 50 states and Washington, D.C., and were reported to the CDC from 2012 to 2013. The 2012 data were included because a county's previous syphilis rates are likely to predict future rates, and they wanted to determine how the Twitter-based method would perform matched with the previous year's data.

They found that counties having higher risk-related tweets in 2012 were associated with a 2.7 percent jump in primary and secondary and a 3.6 percent boost in early latent syphilis cases in 2013. By comparison, counties that reported higher numbers of syphilis cases in 2012 were associated with increases of 0.6 percent and 0.4 percent of primary/secondary and early latent syphilis cases, respectively, in 2013, suggesting that the Twitter-based model performed as well as simply using previous year's syphilis data. This is important because Twitter data are extremely inexpensive and suggest that social media data are low-cost alternatives for predicting syphilis.

Both studies have certain limitations. For the Google paper, they include the likelihood that many primary and secondary syphilis cases are not reported; the findings were biased toward Google users, who account for about 64 percent of search engine users; and the Google Trends data are a random sampling of all data and not the full dataset, which might have affected how the model worked. In the Twitter study's case, data were based on Twitter users, which is a select sample of people; the researchers reviewed data only for 2012 and 2013, when data from a longer time span would be needed to develop appropriate public health responses; and some areas with high numbers of syphilis cases may have had public health messaging via social media that contained relevant keywords that were captured in the data the researchers examined.

More information: Sean D. Young et al, Using Search Engine Data as a Tool to Predict Syphilis, *Epidemiology* (2018). DOI: 10.1097/EDE.0000000000000836

Provided by University of California, Los Angeles