Twitter 'big data' can be used to monitor HIV and drug-related behavior, study shows

February 27 2014, by Enrique Rivero

Real-time social media like Twitter could be used to track HIV incidence and drug-related behaviors with the aim of detecting and potentially preventing outbreaks, a new UCLA-led study shows.

The study, published in the peer-reviewed journal *Preventive Medicine*, suggests it may be possible to predict sexual risk and drug use behaviors by monitoring tweets, mapping where those messages come from and linking them with data on the geographical distribution of HIV cases. The use of various drugs had been associated in previous studies with HIV sexual risk behaviors and transmission of infectious disease.

"Ultimately, these methods suggest that we can use 'big data' from social media for remote monitoring and surveillance of HIV risk behaviors and potential outbreaks," said Sean Young, assistant professor of family medicine at the David Geffen School of Medicine at UCLA and co-director of the Center for Digital Behavior at UCLA.

Founded by Young, the new interdisciplinary center brings together academic researchers and private sector companies to study how social media and mobile technologies can be used to predict and change behavior. (See the center's Twitter account.)

Other studies have examined how Twitter can be used to predict outbreaks of infections like influenza, said Young, who is also a member of the UCLA Center for Behavioral and Addiction Medicine; UCLA's Center for HIV Identification, Prevention and Treatment Services; and

the UCLA AIDS Institute. "But this is the first to suggest that Twitter can be used to predict people's health-related behaviors and as a method for monitoring HIV risk behaviors and drug use," he said.

For the study, researchers collected more than 550 million tweets between May 26 and Dec. 9, 2012, and created an algorithm to find words and phrases in them suggesting drug use or potentially risky behaviors, such as "sex" or "get high." They then plotted those tweets on a map to discover where they originated, running statistical models to see if these were areas where HIV cases had been reported.

The algorithm captured 8,538 tweets indicating sexually risky behavior and 1,342 suggesting stimulant drug use. The geographical data on HIV cases to which researchers linked the tweets came from AIDSVu.org, an interactive online map that illustrates the prevalence of HIV in the U.S.; this mapping data was from 2009.

The states with the largest proportion of geo-located tweets, both general as well as HIV-related, were California (9.4 percent), Texas (9.0 percent), New York (5.7 percent) and Florida (5.4 percent). On a per capita basis, the largest raw number of HIV risk–related tweets came from the District of Columbia, Delaware, Louisiana and South Carolina. States with the highest per capita rate of tweets were Utah, North Dakota and Nevada.

When the researchers linked the tweets to data on HIV cases, they found a significant relationship between those indicating risky behavior and counties where the highest numbers of HIV cases were reported.

Based on this study, the researchers conclude that it is possible to collect "big data" on real-time social media like Twitter about sexual and drug use behaviors, create a map of where the tweets are occur and use this information to understand and possibly predict where HIV cases and

drug use occur.

The study's main weakness, the researchers say, is that the HIV data comes from 2009, so in order to test if this approach can be used to predict future behaviors and outbreaks there is a need for a "gold standard" of frequently updated data. In this way, tweets can be accessed instantly to compare them with disease outbreaks.

The study does however demonstrate the feasibility of using real-time social networking to identify and map HIV risk-related communications and link them to national HIV data, the researchers write.

"This study was designed to call for future research to understand the potential cost-effectiveness of this approach and to refine methods of using real-time social networking data for HIV and public health prevention and detection," they conclude.

Provided by University of California, Los Angeles