The Next Frontier: Digital Disease Detection in Cyberspace

It is understood that any country's innovation pipeline, competitive edge, and domestic productivity will increasingly depend on the available deep analytic talent. Trillions of data are being captured day-by-day by governments, companies, and the scientific community. Sensors embedded in digital devices monitor a host of parameters that, combined with the digital capture of individual actions and behaviors, including the chatter in social networks, offer new venues for research on huge data sets. Although much effort is focusing on consumer behavior, questions about health and well-being are not excluded from this new trend of mining large observational data sets.

Unlike waiting for the US Centers for Disease Control and Prevention (CDC) or some other national entity to receive messages from clinicians and diagnostic laboratories about the occurrence of newly diagnosed cases in support of declaring an epidemic, there is the assumption that the existence, or even the likely occurrence, of such can be predicted based on mining the chatter on social networks. What has become known as "computational social science" aims to produce tools to impact positively the well-being in humans. Using the Internet for identifying disease outbreaks, health surveillance, and other types of health intelligence by means of scanning blogs, chat rooms, social networks, or entries on web search engines has become a new source for what can be called "digital disease detection." Who would have guessed that unstructured but somewhat relevant bytes of information, buried in big data sets that are transmitted over the Internet and are orderly mapped geographically, would carry information of potentially crucial relevance to global public health? The idea that unstructured data in digital networks could become a tool for global disease preparedness is mind-boggling for one who first used the Internet by placing the handset of a phone on an acoustic modem. In recognizing the opportunities for mapping the occurrence and spread of infectious disease as a logical first investment, the notion then to link geographically mapped epigenetic factors to the occurrence of chronic disease and associated symptomatic, comorbid conditions no longer appears to be as unrealistic.

Take the Google Flu Trends as an example. The computational algorithms used to predict flu epidemics are indeed temporally congruent with the traditionally measured fluctuations in case incidence, but they may either over- or underestimate the penetration of disease in a given population. The algorithms that are employed to arrive at these measures of disease penetration within or beyond a given population require recalibration from time to time to predict values comparable to those derived from traditional, needless to say, much more cumbersome approaches. It is assumed that a strong, concurrent media focus may result in the overestimation of disease.

As investigations are digging deeper and deeper into the "omics" associated with a common human disease phenotype, we learn that often less than half of the phenotype is explained by available "omic" data. Searching for the remaining portion of the unexplained variance may require investigators to look at the digital fingerprint that diseases leave behind in big data to determine whether geo-positioned and co-located risk factors play a role in accounting for the remaining unexplained variance in disease risk.

Who would have guessed 20 years ago that a disease would leave an imprint in the chatter traveling in fiber optic cables that could help traditional scientific inquiries by means of mining the sensors that are built into the "toys" and the way we use them. This new frontier has much to offer.

Christian Stohler
Associate Editor