Mining the data mother lode

June 7 2017, by Karen Kreeger

Credit: Sean Ellis

A mother posting on Facebook about the way her son behaves while playing video games could provide a vital clue for the correct treatment for his epilepsy. This is but one type of social media chatter that is informing data scientists at Penn Medicine's Health Language Processing Lab (HLP). One of the newest entities with the Penn Institute for

Biomedical Informatics, HLP combines social media content with other sources of health information in a unique way aimed at understanding how people use language to communicate health needs.

All sorts of groups across Penn Medicine are harnessing data contained in electronic health records (EHRs) and social media to help improve outcomes. The Abramson Cancer Center, for instance, uses lab tests, radiology visits, and patient-reported symptoms to help lung cancer patients avoid the ER visits. Departments across the health system work with data scientists to use finely tuned algorithms to detect complications or underlying health conditions earlier in the continuum of care, and the Center for Digital Health focuses its efforts on how social media intersects with health care, working to determine how posts might help providers detect health problems before urgent care is needed, or even how Twitter might play a role in fighting HIV.

"As other initiatives focus on structured data, the Health Language Processing Lab takes another approach. We're not only about social media or data gained from electronic medical records," said Graciela Gonzalez-Hernandez, PhD, an associate professor of Biostatistics and Epidemiology and director of the HLP. "We also use natural language processing techniques on different sources of information—health records, patient reports, consumer reviews, clinical records, and published literature. Then we use data science techniques to integrate them and present them to specialists for context and discovery."

The HLP uses similar tools as other Penn bioinformatics studies. For example, PennSeek (think Google) is mining unstructured data in clinical records, such as handwritten notes, to refine patient care in cardiology and ophthalmology. Gonzalez-Hernandez came to the field of bioinformatics by way of an undergraduate degree in journalism. She brings her love of words to her present endeavors. "Health tweets have a special twist compared to other social media chatter," she said. "We see

more metaphors, similes, and even sarcasm. This source is rich with meaningful content that can be used to gain insights into the health not only of the person posting, but of a group of people akin to the poster."

One of HLP's projects, sponsored by AbbVie Pharmaceuticals, is focused on improving knowledge about the use and effects of drugs and vaccines during pregnancy, with the long-term goal of finding associations between medication used and fetal outcomes. Currently, all information gathered on this topic is through pregnancy registries. Since these health system- and industry-sponsored databases are voluntary, they have limitations, such as low enrollment rate (a majority of pregnant women that are taking medications do not register), high cost, and selection bias (most information is entered only after something bad happens). "We are assessing Twitter to see if we can broadly monitor health information for large groups of pregnant women who take different kinds of medication, such as over-the-counter pain relievers," Gonzalez-Hernandez said.

Her group also uses natural language processing to mine the clinical records of pediatric epilepsy patients to detect patterns of symptoms. "Epilepsy is a very complex disease and we're trying to find the right treatment for the right patient," Gonzalez-Hernandez said. "For example, parents might share that their son was playing a video game, describing his behavior as 'he sat there frozen' and another parent might describe their child's behavior as 'zoning out.'" Each of these states represents different types of seizures, and recognizing those differences is crucial to choosing the right treatment.

All of these projects are promising for the future of patient care, but Big Data analysis will need quality control like any other area of patient care. "A new approach called continuous analysis will shake up biomedical data science by reducing many common types of analytical errors," said Casey Greene, PhD, an assistant professor of Pharmacology. He likens

continuous analysis to the movie Groundhog Day: "Phil Connors [Bill Murray's character] has to live the same day over and over again. Each time he gets things a little better until he gets it exactly right. This is how reproducible data science should work."

Greene explains that each analysis should be repeated from start to finish with each small change to get things right. For example, data scientists might swap statistical tests when a variable doesn't meet the assumptions of the test they planned to use. Then they would repeat all the steps from start to finish.

"But, this isn't exactly how things are done right now," Greene said. "It's time-consuming for a human data scientist to redo things from start to finish. Remember how frustrated Phil got going through the motions over and over again." Data scientists take shortcuts, he says. They go back to what they think is the right point, and work from there. But this opens up the door to mistakes.

Instead, with continuous analysis, a program watches for any changes to the workflow. When a data scientist makes changes, it automatically runs every step from start to finish. The results are produced and permanently recorded so that anyone can go back and see what the plan was and what the results of that particular "Groundhog Day" day were.

The purpose is to make it easy to build on what other data scientists have done. "Science is incremental," said Greene. "Brick by brick, we're building scientific knowledge."

Provided by University of Pennsylvania