MedicalXpress

Mining mountains of data for medical insights

June 24 2014, by Michael Haederle

Epidemiologists know that an important piece of evidence is often staring you in the face – but it's not always easy to see the forest for the trees.

Danish scientists recently teamed up with University of New Mexico researchers to test a powerful new method for predicting the progress of common diseases through time by teasing out previously undetected patterns from a very large data set – in this case, the health records of Denmark's entire population.

This approach maps out surprising correlations: a disease like gout – a form of arthritis – is strongly linked to cardiovascular disease, for example. In the future, this could enable physicians to make diagnoses sooner using simple tests in combination with known disease progression patterns.

The research is outlined in a study, "Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients," published Tuesday in *Nature Communications*.

Pope Moseley, MD, chair of UNM's Department of Internal Medicine and Tudor Oprea, MD, PhD, professor of Internal Medicine and chief of UNM's Translational Informatics Division, collaborated with researchers from the Department of Systems Biology at the Technical University of Denmark, the Novo Nordisk Center for Protein Research at the University of Copenhagen and the Institute of Biological Psychiatry at

Copenhagen University Hospital.

"This is a leap into a fairly large data base," Moseley says. "This method is able to recognize patterns in data that not only include diagnostic patterns, but includes the element of time and is able to build networks from that."

Denmark's electronic health registry covers that nation's entire population, with each person assigned a health number, Moseley says. Each medical diagnosis is coded in the registry using the International Classification of Diseases terminology – 101 million unique diagnoses in all.

"Every diagnosis on every Dane from every hospitalization and outpatient clinic visit is entered into the national health registry for the last 14 years," he says. "You're able to take these mass of data and look at it over time and begin to draw associations."

The team boiled down the massive trove of data to 1,171 so-called thoroughfares with central information on the course of diabetes, chronic obstructive pulmonary disease (COPD), cancer, arthritis and cardiovascular disease.

Lead author Anders Boeck Jensen says this data analysis method made it possible to view diseases in a larger context.

"Instead of looking at each disease in isolation, you can talk about a complex system with many different interacting factors," says Jensen, a post-doctoral fellow at the Center for Protein Research. "By looking at the order in which different diseases appear, you can start to draw patterns and see complex correlations outlining the direction for each individual person."

Oprea points out an additional advantage of the data-mining method. "The disease trajectories in this study follow causal relationships that were identified by a medically agnostic software," he says. "This illustrates the power of data mining as a means to uncover novel disease relationships and its ability to inform the health care sector about new avenues in patient management."

The data analysis showed, for example, that a diagnosis of anemia is typically followed months later by the discovery of colon cancer, Oprea says, "which suggests that cancer lesions were present and occult bleeding occurred, but remained undiagnosed."

Meanwhile, in addition to identifying gout as a step on the path toward cardiovascular disease, the team made surprising findings about COPD.

"In just looking at these codes that were based on age and gender and where the code was done, we were able to say that COPD is diagnosed late," Moseley says. "It's therefore under-diagnosed and probably because of that undertreated. All we have is this diagnostic code, but our analysis of the pattern said that."

That finding received unexpected support last February when another team published a paper on a large epidemiological study of 6,000 Danish COPD patients, each of whom was interviewed and subjected to extensive examination, laboratory review and testing.

"Their conclusion is COPD is diagnosed late, under-diagnosed and undertreated," Moseley says. "We were able to come to the same conclusions without ever having gone the other way. We essentially did the experiment with a computer out of a health registry."

The research could yield tangible health benefits as we move beyond one-size-fits-all medicine, says Prof. Lars Juhl Jensen of the Center for

Protein Research.

"The perspective is that your genetic profile or the total network of associated proteins in your body, your proteome, can be mapped in a few years' time, enabling you to suddenly learn things about yourself which can be used to forecast the progress of diseases over an entire lifetime," he says.

Søren Brunak, a professor at the Technical University of Denmark and Center for Protein Research who served as senior author on the paper, adds that the sooner a health risk pattern is identified, "the better we can prevent and treat critical diseases."

Moseley describes the partnership with the Danes as "a really very nice marriage . . . it's a strong informatics and systems biology collaboration." Going forward, he hopes to access the data for even larger populations.

"The author Williams Gibson said something like, 'Everything we need to know about the future is here, now – you just have to be able to recognize the pattern,'" Moseley says. "Never was it more true."

More information: *Nature Communications*, dx.doi.org/10.1038/ncomms5022

Provided by University of New Mexico

provided for information purposes only.