How a computer can help your doctor better diagnose cancer

April 23 2015, by Adam Conner-Simons

Correctly diagnosing a person with cancer—and identifying the specific type of cancer—makes all the difference in successfully treating a patient.

Today your doctor might draw from a dozen or so similar cases and a big book of guidelines. But what if he or she could instead plug your test results and medical history into a computer program that has crunched millions of pieces of similar data?

That sort of future is looking increasingly possible thanks to researchers at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL). Working with a team from Massachusetts General Hospital (MGH), PhD student Yuan Luo and MIT Professor Peter Szolovits have developed a computational model that aims to automatically suggest cancer diagnoses by learning from thousands of data points from past pathology reports. The work has been published this month in the *Journal of the American Medical Informatics Association.*

Better lymphoma diagnoses

The researchers focused on the three most prevalent subtypes of lymphoma, a common cancer with more than 50 distinct subtypes that are often difficult to distinguish. According to Ephraim Hochberg, director of the Center for Lymphoma at MGH and one of the paper's co-authors, upwards of 5 to 15 percent of lymphoma cases are initially misdiagnosed or misclassified, which can be a significant problem since different lymphomas require dramatically different treatment plans.

For example, Hochberg recently saw a patient who had been mistakenly told that her lymphoma was incurable. If he hadn't accurately diagnosed her and put her on an aggressive plan, it might have been too late to counteract the cancer.

Lymphoma classification has long been a source of debate for pathologists and clinicians. There were at least five different sets of guidelines until 2001, when the World Health Organization (WHO) published a consensus classification. In 2008 the WHO revised its

guidelines in a labor-intensive process that involved an eight-member steering committee and over 130 pathologists and hematologists around the world. In addition, only around 1,400 cases from Europe and North America were reviewed to cover 50 subtypes, meaning that on average a subtype's diagnosis criteria was based on what happened to only a limited number of people.

Meanwhile, large medical institutions like MGH often archive decades of pathology reports. This got the MIT researchers thinking about whether they could tap into these resources to develop automated tools that could improve doctors' understanding of how to diagnose lymphomas.

"It is important to ensure that classification guidelines are up-to-date and accurately summarized from a large number of patient cases," says Luo, who is first author on the paper. "Our work combs through detailed medical cases to help doctors more comprehensively capture the subtle distinctions between lymphomas."

Doctor-friendly models

Luo emphasizes that such machine-learning models need to be not only accurate but also interpretable to clinicians. The WHO guidelines' criteria are outlined via a panel of test results that are themselves relations among medical concepts such as tumor cells and surface antigens. In order to capture the relations, the researchers converted sentences from pathology reports into a graph representation where graph nodes are medical concepts and graph edges are syntactic/semantic dependencies. As described in their previous paper, they then collected frequently occurring subgraphs that correspond to relations that specify test results.

"Clinicians' diagnostic reasoning is based on multiple test results

simultaneously," Luo says. "Thus it is necessary for us to automatically group subgraphs in a way that corresponds to the panel of test results. This makes the model interpretable to clinicians instead of being a black-box, as they often complain about many other machine learning models."

The core contribution of this work is to use a technique called Subgraph Augmented Non-negative Tensor Factorization (SANTF). In SANTF, data from the 800 or so medical cases are organized as a three-dimensional table where the dimensions correspond to the set of patients, the set of frequent subgraphs, and the collection of words appearing in and near each data element mentioned in the reports. This scheme clusters each of these dimensions simultaneously, using the relationships in each dimension to constrain those in the others. By examining the resulting clusters, the researchers can link test result panels to lymphoma subtypes.

"The promise of Luo's work, if applied to very large data sets, is that the criteria that would then help to identify these clusters can inform doctors about how to understand the range of lymphomas and their clinical relationships to each other," Peter Szolovits says.

"Most natural-language processing in clinical reporting has focused on identifying important phrases or attributes, and not the more difficult task of recognizing relationships and concepts," explains Professor Wendy Chapman, chair of the department of biomedical informatics at the University of Utah. "Medical experts with years of experience are able to understand not just the words, but the deeper implications. This research gets us a step closer to developing robust computer models that can achieve that level of comprehension."

On top of that, the SANTF model does not require labeled training data, which makes it possible to automate the process of knowledge discovery. Szolovits is confident that that the team's model can help doctors make

more accurate lymphoma diagnoses based on more comprehensive evidence—and could even be incorporated into future WHO guidelines.

"Our ultimate goal is to be able to focus these techniques on extremely large amounts of lymphoma data, on the order of millions of cases," says Szolovits. "If we can do that, and identify the features that are specific to different subtypes, then we'd go a long way towards making doctors' jobs easier—and, maybe, patients' lives longer."

More information: "Subgraph Augmented Non-Negative Tensor Factorization (SANTF) for Modeling Clinical Narrative Text" DOI: dx.doi.org/10.1093/jamia/ocv016

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology