BY JONATHAN WOSEN

DISEASE DETECTIVES:
BD2K CENTER RESEARCHERS SLEUTH FOR EARLY SIGNS OF DISEASE

Most people don't know that they're sick until they feel, for lack of a better word, *sick*. Like storms, diseases quietly brew and gather strength before wreaking havoc. For the weather, however, you can turn on your local news channel and check next week's forecast. Not so for disease. Not yet, anyway. Researchers across the NIH Big Data to Knowledge (BD2K) Centers are pursuing innovative, data-driven strategies to predict disease and its progression.

Such predictions would help doctors and scientists alike, says **Mark Craven**, **PhD**, professor of biostatistics and medical informatics at the University of Wisconsin-Madison (UW-Madison) and director of the BD2K **Center for Predictive Computational Phenotyping** (**CPCP**).

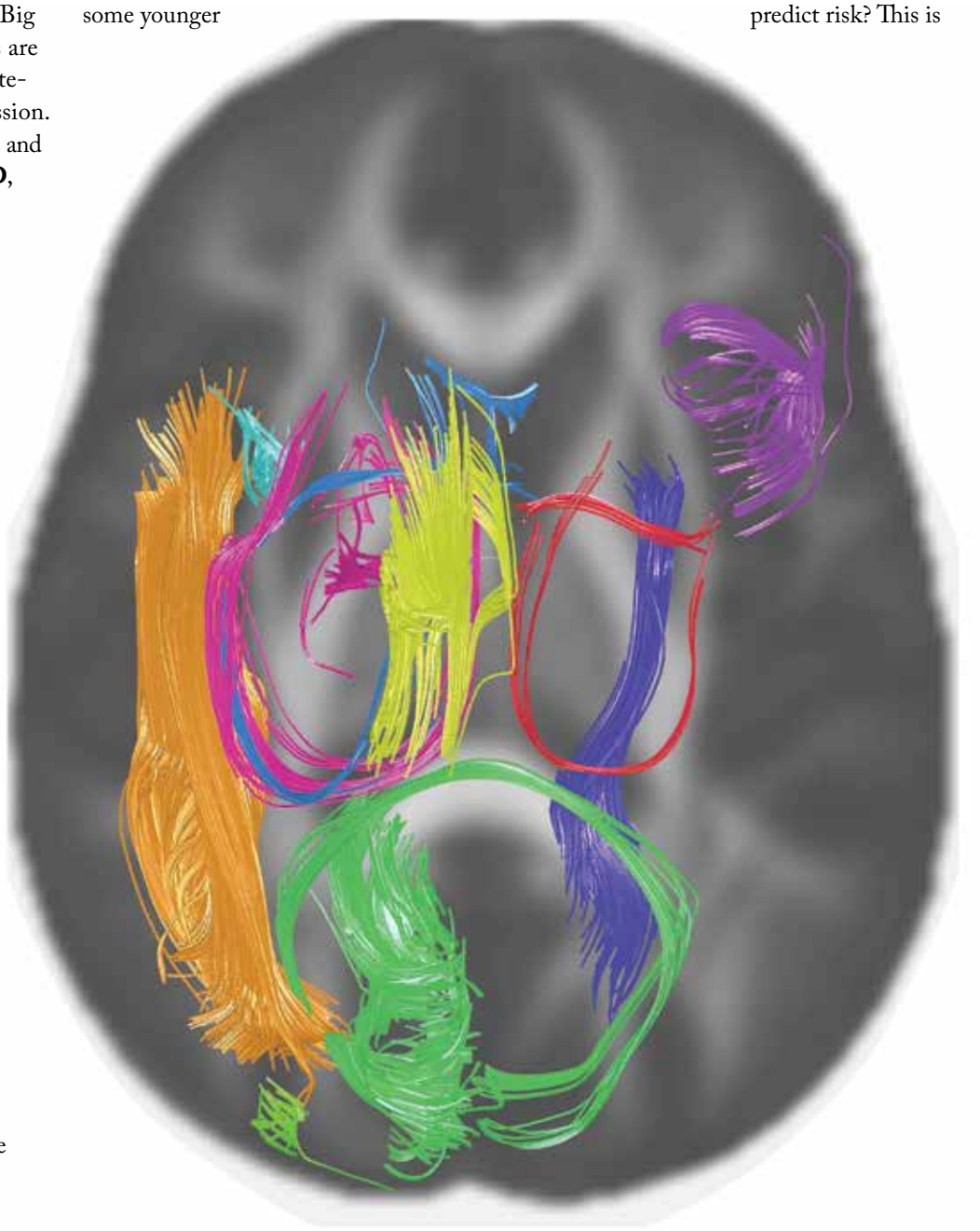
For many conditions, if you can predict that it's headed your way, Craven says, "that can give clinicians some kind of guidance."

Armed with predictive models, doctors could intervene sooner to improve patient outcomes. Statistical models could also predict whether a patient's disease will progress quickly, slowly, or hardly at all. Identifying who falls into which group may be the key to choosing candidates for clinical trials.

BD2K researchers are using their computational toolkits to study everything from neural changes that presage Alzheimer's disease to rates of osteoarthritis progression. Here, we feature a few stories of exceptional disease detective work with the potential to reshape our understanding of when and why diseases strike.

Detecting Disease Earlier

For many diseases, there are official guidelines regarding screening patients. Take breast cancer, for instance. The American Cancer Society recommends women age 45 and up receive regular mammograms. But some younger women are at higher risk than women in older age groups. Should a woman in her early 40s with a family history of breast cancer be screened? A doctor would have to consider family history, demographics, and a patient's medical record to come up with an answer. So why not have an algorithm help predict risk? This is



RELEVANT NIH INSTITUTES:

NCI, NHLBI, NIAMS, NIA, NINDS, NCATS, and all other disease-focused Institutes

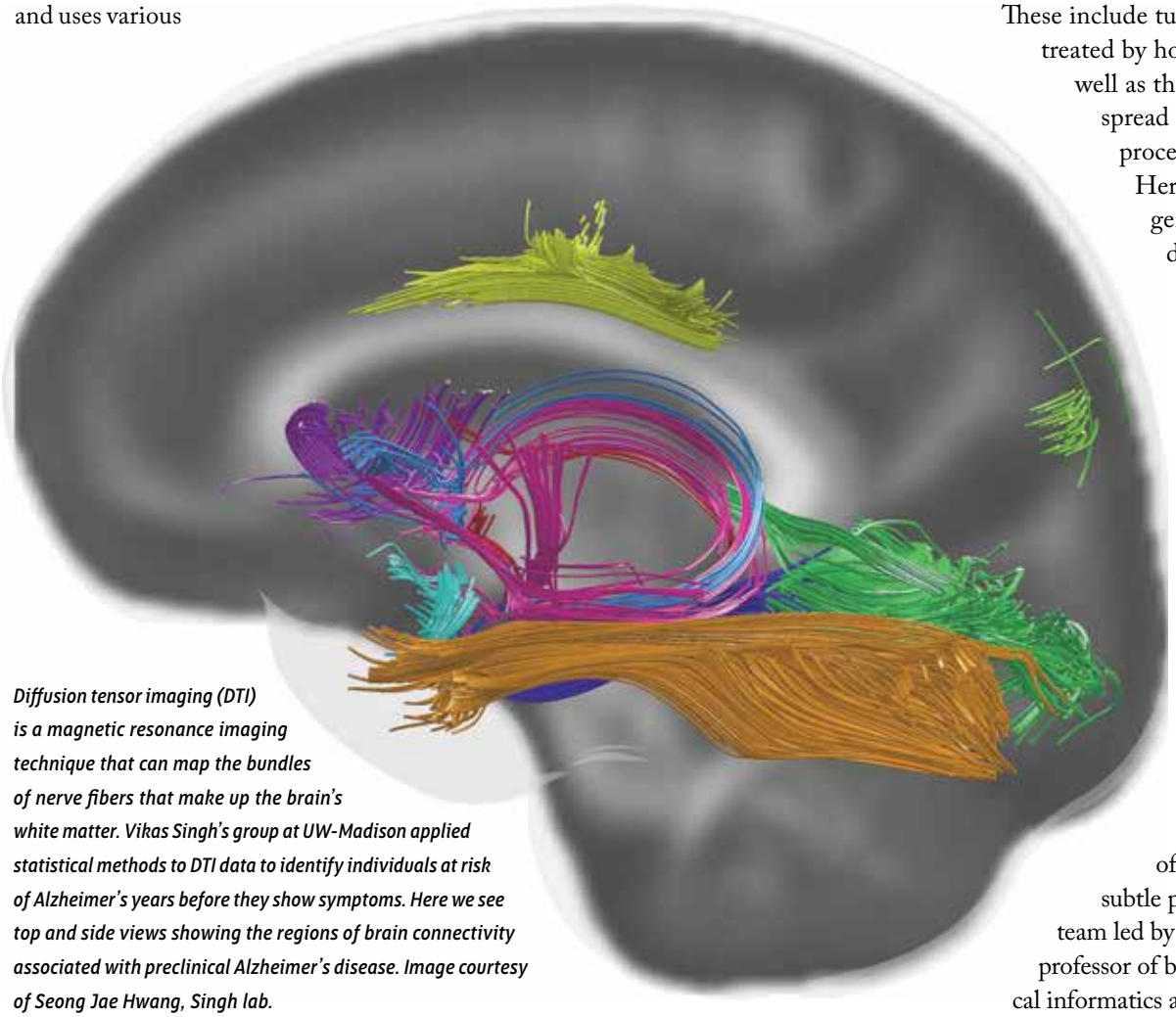
the goal of **Elizabeth Burnside, MD**, professor of radiology at UW-Madison.

"What we envision is … a tailored approach depending on a woman's risk and values," Burnside says. "That would hopefully result in better outcomes."

Burnside's team has access to nearly 70,000 mammograms collected at UW-Madison's hospital dating back to 2006, as well as genetic data and personal risk factors (e.g., age, family history, etc.) drawn from those patients' electronic health records (EHRs). In collaboration with CPCP investigators, Burnside's group combines these data and uses various machine learning approaches, including support vector machines, neural networks, and deep learning, to predict breast cancer risk.

Breast cancer is caused by genetics and environmental factors that affect estrogen levels, ranging from diet and exercise to breastfeeding history. Burnside believes it is essential to incorporate both nature and nurture into effective, user-friendly models that will be useful in clinical settings. "If a patient and physician are going to use a model, they generally want to understand how it's working," she says.

"If a patient and physician are going to use a model, they generally want to understand how it's working," Burnside says.



Diffusion tensor imaging (DTI) is a magnetic resonance imaging technique that can map the bundles of nerve fibers that make up the brain's white matter. Vikas Singh's group at UW-Madison applied statistical methods to DTI data to identify individuals at risk of Alzheimer's years before they show symptoms. Here we see top and side views showing the regions of brain connectivity associated with preclinical Alzheimer's disease. Image courtesy of Seong Jae Hwang, Singh lab.

Burnside especially wants increased screening for women at risk of developing aggressive forms of breast cancer. These include tumors that cannot be treated by hormone therapy, as well as those that break off and spread throughout the body, a process known as metastasis. Her group is analyzing genetic and imaging data to determine what groups of women are at risk for aggressive breast cancer so that they can be screened more intensively.

"What we're trying to do is to intervene at the right time in the right patient to decrease the chance of poor outcomes," says Burnside.

Other groups are also capitalizing on the power of imaging to detect subtle phenotypes. A CPCP team led by **Vikas Singh, PhD**, professor of biostatistics and medical informatics at UW-Madison, is

developing statistical algorithms that use diffusion MRI data to map degeneration in brain connections in patients at risk of developing Alzheimer's.

While memory loss and confusion are hallmark Alzheimer's symptoms, they are preceded by a decades-long preclinical phase of the disease. Singh and his collaborators wanted to understand how the brain's intricate web of neural connections change during this early phase of the disease.

"Once you are able to identify or predict this future disease course, then you can identify which subset of individuals are most likely to be helped by a new treatment," says Singh.

Singh's team, including graduate students **Won Hwa Kim** and **Seong Jae Hwang** and research scientist **Nagesh Adluru**, **PhD**, analyzed MRI data collected with diffusion tensor imaging, which uses the diffusion of water molecules to reveal tissue architecture. The data, collected at the Wisconsin Alzheimer's Disease Research Center in studies led by UW-Madison professors **Sterling Johnson**, **PhD**, and **Barbara Bendlin**, **PhD**, revealed a variety of neural connections that differed in strength between cognitively normal adults with and without a first-degree relative with Alzheimer's. Previous studies have shown that individuals with a family history of Alzheimer's are more likely to develop the disease.

Further research on these connections and the brain regions they encompass may shed light on how Alzheimer's progresses. Singh and his collaborators are now investigating how structural connectivity changes correlate with known protein biomarkers of Alzheimer's, such as beta-amyloid and tau.

Both Burnside and Singh have focused their analyses on specific diseases, with the goal of applying their methods to other conditions. **David Page**, **PhD**, professor of biostatistics and medical informatics at UW-Madison, takes a distinctly broader approach.

"We have lots of EHR data. How well can we predict *every* diagnosis that a patient is going to get?" Page says.

Using 40 years of de-identified EHR data from 1.5 million patients at the Marshfield Clinic in north-central Wisconsin, Page's team built predictive models for nearly 4,000 diseases. Their strategy used random forests, a classification algorithm that uses decision trees to guide predictions, and required HT-condor, a high-throughput computing environment that could handle the deluge of data.

Using 40 years of de-identified EHR data from 1.5 million patients at the Marshfield Clinic in north-central Wisconsin, Page's team built predictive models for nearly 4,000 diseases.

The researchers predicted disease from one month to 20 years in advance. All predictions were better than random chance, though, as expected, earlier predictions were less accurate. Page believes that, with further research and improved model accuracy, his 'pan-diagnostic' approach can become a widely used tool that supports both providers and patients.

"We'd like to explore whether some of these models are good enough … to translate them into use in the clinic—and test whether that has a positive impact," Page says.

Forecasting Disease Progression and Complications

Once patients learn that they have a disease, they want to know how it will progress. Will their symptoms steadily worsen, plateau, or alternate between active and inactive periods? These questions also matter to doctors as they decide the best course of treatment.

At the **Mobilize Center** at Stanford, **Eni Halilaj**, **PhD**, postdoctoral fellow in the lab of **Scott Delp**, **PhD**, is studying the progression of osteoarthritis. Halilaj and her Stanford collaborators are analyzing X-rays from the Osteoarthritis Initiative, a multi-center study of knee osteoarthritis progression. Osteoarthritis wears away cartilage over time, which appears on an X-ray as a narrowing of the distance between bones.

Halilaj is building a model that combines information from an initial X-ray with dietary habits, medical histories, joint exam and performance measures, and baseline symptoms to identify slow versus fast progressors—a potentially confounding factor in clinical trials.

"The goal is to predict the kind of progressor that someone will be so that we can balance treatment and control groups in…clinical trials," says Halilaj.

The same statistical tools that predict disease progression can be adapted to other adverse clinical

"Once you are able to identify or predict this future disease course, then you can identify which subset of individuals are most likely to be helped by a new treatment," says Singh.

MINING FOR PAIN

BY JONATHAN WOSEN

Over a million patients get joint replacements each year in the United States, often due to osteoarthritis, a leading cause of disability. Approximately five percent of replacements fail, according to the American Academy of Orthopedic Surgeons. And postoperative pain can be an indicator that a device is failing.

"We are interested in asking the question, 'Can we mine electronic health records for device surveillance?'" says **Alison Callahan, PhD,** research scientist in **Nigam Shah**'s biomedical informatics lab at Stanford University. Specifically, she wants to determine whether tracking postoperative pain can provide insight into the effectiveness of specific implant models.

"We are interested in asking the question, 'Can we mine electronic health records for device surveillance?'" says Callahan.

There's just one problem: Most mentions of pain aren't neatly coded in a patient's electronic health record (EHR). Instead, Callahan must dive into the deep, murky waters of unstructured data. Physician notes are a treasure trove of information but are riddled with typos, different ways of referring to pain, negative statements ("the patient did *not* experience pain"), and hypotheticals ("*if* the patient has pain"). It's a job that, in and of itself, can be quite painful.

To help her mine EHRs from Stanford Hospital and Clinics for, as she puts it, "the type of pain someone's having and where it hurts," Callahan needed labelled training data. Using experts to manually label data would be expensive and time-consuming, so she turned to Snorkel, a tool developed by Mobilize Center researchers in **Christopher Ré**'s lab. Snorkel uses a set of rules, or labeling functions, to create large sets of labelled training data. In Callahan's case, these rules include whether a clinical note contains pain-related terms and information about sentence structure to ensure a true mention of pain. In collaboration with several Ré lab members (postdoctoral fellow **Jason Fries, PhD,** graduate student **Alex Ratner** and postdoc **Stephen Bach, PhD**), Callahan used Snorkel to extract mentions of pain and pain location from the notes of roughly 5,000 hip implant patients. She then tested the extraction accuracy with a small subset of data that was manually labeled with the aid of a physician.

Callahan has presented her work at the 2016 Stanford Data Science Initiative retreat, and her initial extraction results look promising. She now plans to scale up to include larger data sets. Because Snorkel is a general system, Callahan says, it can be used for other research questions as well. "There are other types of experiences which a patient might report which would get captured in a clinical report, [such as] activities of daily living," Callahan says. As a result, she says, Snorkel has broad applicability for mining unstructured data without the burden of manually labelling large sets of training data.

events, such as postoperative complications. Complications such as infection, heart attack and stroke are major concerns, and studies show that two of every five patients who experience a complication will have more than one. Mark Craven wants to help hospitals understand and predict chains of postoperative complications, which he likens to a snowball effect.

Craven's team utilized a national database of postoperative outcomes known as the American College of Surgeons National Surgical Quality Improvement Program. The researchers considered over 20 different postoperative complications, including infection, heart failure, and extended use of a ventilator. They used Markov chains, which model changes between states, to predict the complications that occurred each day over a 30-day period post-surgery.

Mark Craven wants to help hospitals understand and predict chains of postoperative complications, which he likens to a snowball effect.

The models were particularly accurate for major complications such as death, heart attack, and kidney failure. Going forward, Craven plans to incorporate additional clinical information from the dataset to make earlier and broader predictions about patient outcomes. "At the time of surgery, how much risk do I think this patient has for having any complications, specific complications, or multiple complications?" Craven says.

His lab has already developed an accurate predictive model for post-hospitalization blood clots using information from EHRs. Craven plans to test this model in the clinic through a shadow trial—a process of predicting and measuring outcomes without intervening. Predictions will be made about the risk of clots in specific patients as they are monitored over time. If the predictions hold true, doctors may one day use Craven's model to determine who should be given a blood thinner to prevent clotting after hospitalization.

To Causality and Beyond

Ultimately, predictions for disease progression, outcomes and complications will be more accurate when scientists and doctors understand *why* these events happen. Understanding causation would help researchers design specific therapies that target factors directly involved in disease. **Panayiotis (Takis) Benos**, **PhD**, professor and vice chair of computational and systems biology at the University of Pittsburgh School of Medicine and project leader for the BD2K **Center for Causal Discovery** (**CCD**), wants to develop a causal understanding of chronic lung disease to guide treatment design.

Benos and collaborators are analyzing gene expression and other molecular data together with clinical and histology data from the tissues of patients with idiopathic pulmonary fibrosis. His team uses probabilistic mixed graphical models (MGMs) to combine different data types into a network that reveals direct, causal connections between variables. Using data from the Lung Genomics Research Consortium and new data generated by Benos' team, the researchers have also built MGMs for chronic obstructive pulmonary disease. These models provide insight into how these chronic lung diseases progress and which factors affect the long-term decline of lung function. Knowing these factors, scientists can predict which patients are likely to worsen over the next two to five years, Benos says.

One inherent challenge with this approach is dealing with variables that *aren't* measured. An MGM may show that a certain gene or measurement is directly associated with a disease, but there could be another untested variable that is in closer association. To bolster his models of lung disease, Benos plans to include larger patient cohorts and incorporate additional variables, including CT scans, biomarkers, and patient symptom questionnaires. In addition, his group is developing algorithms to detect when two variables are controlled by an unmeasured lurking variable. This can help scientists and clinicians recognize when they need to collect additional data. Benos believes this graphical approach can reveal causal relationships in other illnesses, and wants to share his team's analytical tools with the scientific community. "We are planning to apply [MGMs] to cancer, influenza and pneumonia. We also plan to have an R package out soon, so people can easily incorporate our method into their own analysis," he says.

BD2K Synergies

The BD2K Centers' contribution to the prediction of disease and its progression is still expanding. To provide a fuller picture of changes in a patient's health between doctor's visits, Page would like to supplement EHRs with data from wearable devices that track blood pressure, heart rate, and body temperature. Research out of the **Mobilize Center** and **MD2K** (**Mobile Sensor Data-to-Knowledge**) could potentially help with that (see "Mobile Health: BD2K Centers Harness Sensor Data," page 10).

In addition, predictive algorithms will be more accurate when built using larger data sets from patients at multiple research centers, which raises the question of how to efficiently share data across centers while also protecting patient privacy. Work out of several BD2K Centers will surely make that a lot easier as well (see "The FAIR Data-Sharing Movement: BD2K Centers Make Headway," page 33).

BD2K has fostered great interactions, Page says. "There's a natural synergy. There's a lot of teamwork." □