

Physicians are forever recording information about their patients. They take vital signs, order lab tests and imaging, prescribe medications, check boxes to define patients' diagnoses for billing purposes, and write or dictate narrative descriptions of each patient's status. For the most part, all of

this information goes into the patient's electronic health record (EHR) where it remains untouched for any purpose other than billing or the patient's next visit.

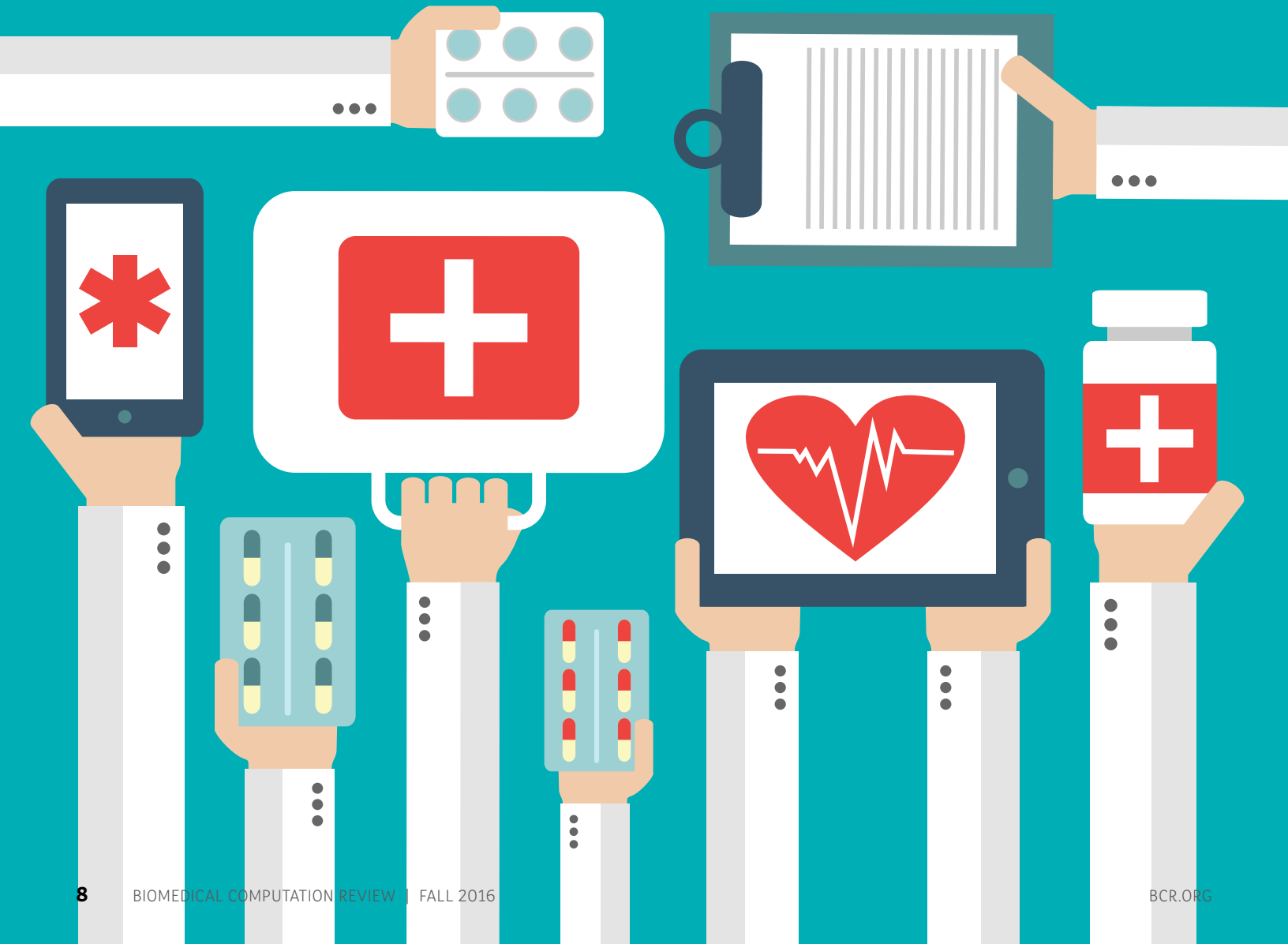
These EHRs represent a vast untapped gold mine for improving patient care. "There is no other industry that doesn't learn from its prior customers," says

Nigam Shah, MBBS, PhD, associate professor of medicine at Stanford University.

In clinical settings, EHRs can be mined to identify patients at high, medium, and low risk for various outcomes, allowing healthcare providers to intervene proactively. For example: Who is likely to be admitted to the ICU or ER?

LEARNING FROM Patients' Health Records: *Bringing Machine Learning to the Clinic*

BY KATHARINE MILLER



EHRs can also be used to personalize risk assessment. For example, someday, a clinician might be able to query a warehouse of EHR data to find how other patients highly similar to one of theirs fared when given various treatments.

EHRs can also be used to predict differences in how diseases progress. For example: When will pre-diabetes progress to full-onset diabetes? Or when will an aplastic mole progress to full-blown melanoma?

Applying machine learning to EHRs for the benefit of patients has its challenges. Medical record systems vary among institutions, are not standardized, and are constantly evolving; diagnostic codes used for billing purposes are often unreliable; and narrative descriptions in natural language are hard for computers to interpret. Moreover, privacy concerns limit access to EHRs; datasets from some institutions may be too small to be useful, especially for rare diseases; and when datasets are larger, the statistical challenges exceed an individual clinician's grasp.

There are also methodological hurdles to cross. "There are probably a dozen widely used machine-learning algorithms and thousands of variations," says **David Page, PhD**, professor of biostatistics and medical informatics at the University of Wisconsin's School of Medicine and Public Health. "We try to be very open-minded about what method would work best."

And then there are the economics of it. Institutions like Stanford University, Shah's employer, may be willing to foot the bill for a data warehouse full of EHRs without concern for the financial return, but the larger healthcare industry would have to pay for EHR work using patient-care dollars—and would need to show benefit to specific patients to collect those funds. "We haven't figured that out yet," Shah says. "How do we demonstrate a return on investment when the people who stand to benefit have no skin in the game?"

Despite the challenges, researchers can point to a number of promising projects that are either already benefiting patients or soon will be. "It's phenomenal to see the work get to this point," says

Jenna Wiens, PhD, assistant professor of computer science and engineering at the University of Michigan. "We talk all the time about leveraging EHR data to produce actionable knowledge, but in practice it can be really hard to do. I'm really excited to see where it leads."

Improving the EHR to Improve Care

For EHRs, like other databases, garbage in will produce garbage out: If doctors and nurses aren't entering data accurately, or aren't keeping the records up-to-date, patient care could suffer. Moreover, narrative notes in EHRs often



hide information that could be useful if it were more structured. So some researchers are using machine learning to improve the accuracy and structure of the EHR—which in turn makes the EHR more valuable for machine learning. It's a great way to tackle some low-hanging fruit, says **David Sontag, PhD**, assistant professor of computer science and data science at New York University.

About seven years ago, Sontag, a specialist in machine learning, began working with **Steven Horng, MD**, associate director in the division of emergency informatics at Beth Israel Deaconess Medical Center in Boston, Massachusetts. They wondered if machine learning could be used to structure the patient's chief complaint as it is entered in the EHR by emergency room (ER) triage nurses. The chief complaint is typically a brief, free-text summary of the patient's condition. For example, it might be "chest pain," "hit by car," "pneumonia," or "uncontrolled

bleeding." It is often the first thing the ER physician sees. This important information could be valuable to record in a structured form, but a drop-down menu of chief complaints would be very long and require too much time from nurses in a hurry.

So, using data for 200,000 patients who had been to the ER in the past, Sontag and Horng, along with Sontag's PhD students **Yacine Jernite** and **Yoni Halpern**, trained a machine-learning algorithm to identify what the chief complaint should be for new patients. Implementing the algorithm in an ER setting required that nurses write a 20- to 40-word triage assessment of the patient, in addition to taking vital signs. The machine-learning

algorithm then uses that information to predict and auto-complete a structured entry for the chief complaint. The algorithm relies on a clearly defined ontology of many hundreds of possible chief complaints. The system, which has been running live for about three years, is much loved by the nursing staff at Beth Israel Deaconess Medical Center. They complain immediately whenever the system goes down, Sontag says. And the quality of the chief complaints has improved, judging from how rarely the nurses and physicians override the algorithm's chief complaint suggestions, he says. Moreover, with the chief complaint recorded as structured data, it becomes possible to apply more advanced machine-learning approaches to the data—ones that might seek to classify ER patients at highest risk of death, for example.

The approach can be used to improve the structure of EHRs in other contexts as well. For example, Sontag's group used

machine learning to predict what should be added to or removed from the EHR's patient problem list. This list of a patient's current health issues provides valuable contextual information when a patient presents with a new problem, but it is hard to maintain and keep up to date.

"These are simple examples," Sontag says, "And they demonstrate that even the simplest of machine-learning methodologies can have a significant impact on healthcare."

Taking these efforts further, Sontag has a vision to create a foundation for the next generation of EHRs. To be able to deduce a patient's past and present as well as predict the future requires structured information that doesn't exist in current EHRs. So Sontag wants to use machine learning to automatically convert unstructured data into structured data. It's not an easy task. Machine-learning algorithms typically require training data that has been labeled by experts. That's hard to come by in healthcare settings, Sontag says, and it often doesn't transfer well from one institution to another. So Sontag came up with a solution he calls the "anchor and learn framework." It uses prior medical knowledge known to an expert to identify an anchor in the EHR, (e.g., the fact that seeing metformin and multiple HbA1c measurements means someone is a diabetic) and then uses that anchor as a basis for learning. Experts are needed only for determining the best anchors—not for labeling all of the data.

"It's not doing diagnosis," Sontag says. "We're not finding something someone

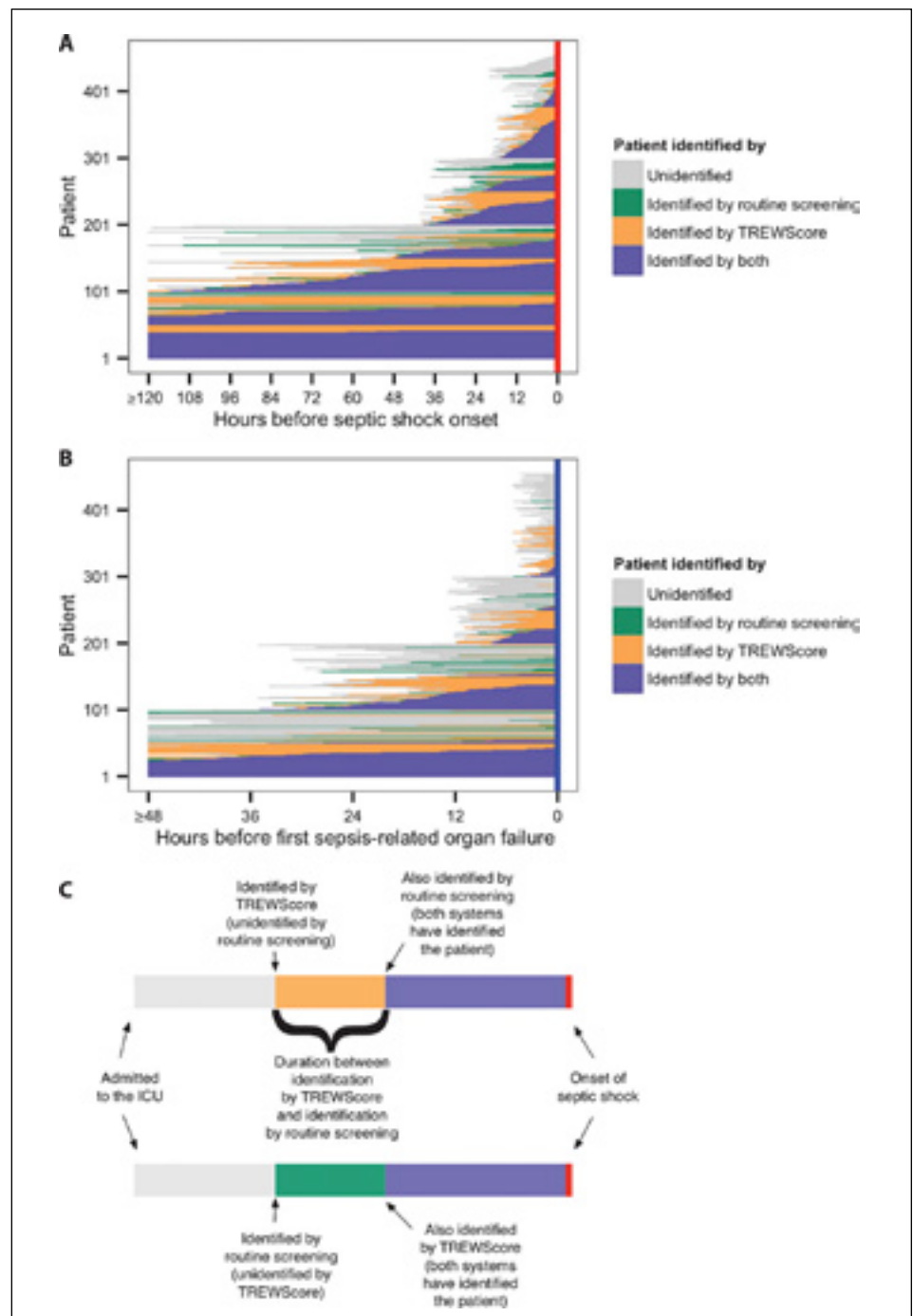
doesn't already know. We're just getting a piece of knowledge that's important into a structured form." For example, if a patient who is being prescribed antibiotics is from a nursing home—a context where antibiotic resistant bacteria often develop—the EHR could flag that and then offer a popup asking "are you sure the patient doesn't have antibiotic resistant bacteria?" But the EHR can only do that if being "from a nursing home" is known. And Sontag's system can figure that out.

Sontag is also looking into using the anchor framework to predict future

events. For example, researchers can look at people who died and then project backward to identify key characteristics in their EHRs several hours or days earlier. These characteristics could then be used as anchors to predict a current patient's likelihood of dying.

Individual Risk Stratification: Predicting Chance of Infection

In hospital settings, patients often face an amplified risk of infection either



Saria and her colleagues compared routine screening procedures to their machine learning-based TREWScore predictions of septic shock during the 120 hour period before septic shock onset (A) and of sepsis-related organ failure during the 48 hours before it occurred (B). Each patient in graphs A & B is represented by a single line (C), with colors reflecting the point at which either routine screening (green) or the TREWScore (orange) or both (purple) predicted septic shock. Thus the quantity of orange in the graphs reflects the success of the TREWScore compared with the quantity of green (routine screening). From KE Henry, DN Hager, PJ Pronovost, S Saria, A targeted real-time early warning score (TREWScore) for septic shock, *Science Translational Medicine* 7:299:122 (2015). Reprinted with permission from AAAS.

because they have an underlying disease, their immune systems are compromised, or they've been overtreated with antibiotics, creating a hospitable environment for antibiotic-resistant bacteria. Predicting which patients are most vulnerable could allow healthcare providers to intervene sooner to prevent or control infections. Already researchers are using EHR data to predict two of the most challenging in-hospital infections: sepsis and *C. difficile*.

Sepsis occurs when the body's response to infection begins to shut down the body's organ systems. It's associated with 20 to 30 percent of all hospital deaths each year in the United States—that's about 750,000 people. Automated screening tools have been used to predict that a patient is experiencing sepsis, but none can predict it in advance. "The question was, 'How can you detect sepsis without having to suspect it?'" said **Suchi Saria, PhD**, assistant professor of computer science at Johns Hopkins University, at the Big Data in Biomedicine Conference at Stanford University. She and her colleagues set out to determine whether EHR-based predictions could outperform the standard of care. They developed a score—the TREWScore—that relies on continuous sampling of the EHR. If the score crosses a certain threshold, it is highly predictive of septic shock.

"Using routinely collected data we were able to predict individuals who experience septic shock on average 25 hours early," Saria said. "That's a huge window for intervention." The work was published in *Science Translational Medicine* in August 2015. Further, she adds, "TREWScore is only a starting point. A lot more can be done to target TREWScore to the individual." Her team is actively working on this and she already sees promise.

Wiens and **Erica Shenoy, MD, PhD**, of Massachusetts General Hospital (MGH) took on a different problem that plagues hospital inpatients: *C. difficile* infection (CDI), which causes diarrhea and colitis. CDI is often caused by antibiotic treatment that eliminates the good bacteria in a person's gut, leaving them

vulnerable to the *C. difficile* bacterium.

Like Saria's sepsis work, Wiens's CDI work generates a score for the probability that a patient will test positive for the infection at a later time during the hospital visit. Her algorithm uses two modeling approaches jointly: a time-invariant predictive model that pools data over several days prior to a positive *C. difficile* test, as well as individual daily models that evaluate which parameters are important on each day leading up to the diagnosis. "Other approaches assume a pattern," she says. "We just let the data speak."

The work, which was published in the *Journal of Machine Learning Research* in 2016, identified both expected and unexpected risk factors that contributed to CDI. Patients taking common antimicrobials or proton pump inhibitors were already known to be at high risk for CDI. More surprising, Wiens says, were factors like location in the hospital and the use of opioids. "It's not clear if that's causal," Wiens says, "But it's a hypothesis that can be tested."

The CDI risk score will be applied next at MGH, and will automatically produce a risk estimate for each patient every day at midnight. Wiens and her colleagues are planning a randomized controlled trial to estimate the potential impact of

risk-driven interventions. The planned study will screen for all patients that are at high risk for CDI, but only intervene in a subset of that group. Wiens and her colleagues will then measure the incidence and severity of CDI

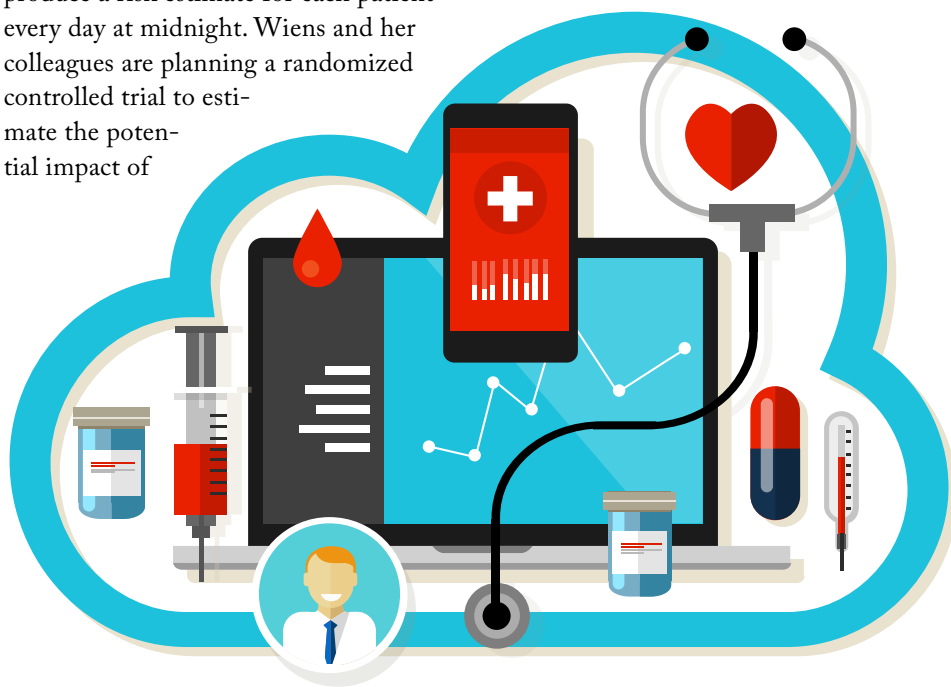
for the two groups and will assess any impact on antimicrobial use and costs.

Modeling each inpatient hospital day and then combining it with a more general model, as Wiens and her colleagues have done for CDI, could prove useful for predicting the progression of other diseases as well. The approach could also generalize more broadly. "You could look at longer time scales to capture how risk factors change over a patient's lifetime," Wiens notes.

Another important direction for the future: combining EHR data with omics data, such as the microbiome. "We're working on that right now," Wiens says. "How much can we predict based on the EHR and microbiome separately versus by combining the two?"

The Informatics Consult: Data Analysis for One Patient at a Time

One of Shah's goals is to develop a medical specialty he calls the "informatics consult." Using machine learning and an EHR warehouse, an informat-



ics expert would be available to advise physicians about the prognosis or treatment options for a particular patient. And clinicians would request a consult just as they do from other medical specialists, such as pathologists or radiologists.

To launch a consult, the clinician would describe the patient—Shah posits a 55-year-old Vietnamese woman with asthma and moderate hypertension—and ask for an appropriate treatment intervention. The clinician knows that an antihypertensive medication is appropriate, but which one works for middle-aged asthmatic females who also happen to be Vietnamese? The informatics consultant would then use the EHR to identify similar patients and the most effective treatments for them. If the EHR system contains only five people who match that patient, the consultant might relax the age or ethnicity conditions to get a bigger sample.

“It makes intuitive sense that being able to make decisions using similar patients would lead to better decisions,” Shah says, “but that’s still a hypothesis.” He plans to test that hypothesis in the coming year. The initial pilot will include a limited number of clinicians who will send a consult request over phone or email. “It’s not fully automated and black-box yet,” Shah says. “People might not trust it; and we’re still not at a stage where, technically, we can shrink wrap it and make it into a button.” But the process would be semi-automated in the sense that the informatics expert gets the question, uses a search engine to find a set of similar patients, and then—depending on the question—applies an appropriate statistical method to the EHR data. “There has to be a human in the loop,” Shah says. But in two to four hours, the consult would generate a predesigned report. “That’s my hope for the first pass,” Shah says.

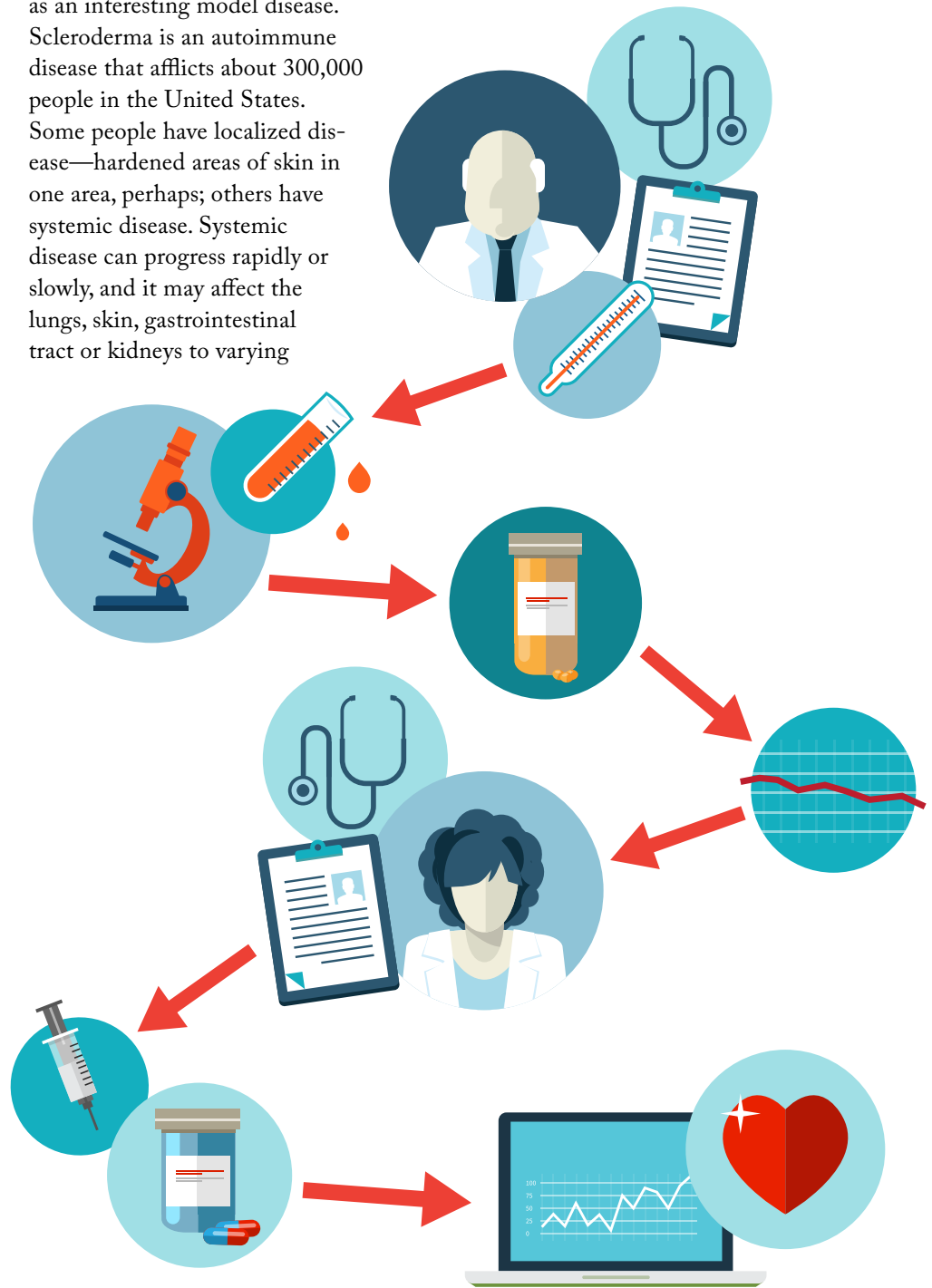
After completing the pilot, they’ll refine the procedure and implement a randomized trial. Some physicians will have access to the consult and others won’t. After a year, Shah’s team will look for differences in outcomes such as the cost of care; speed of recovery; and patient well-being and satisfaction

Predicting Disease Progression

One of the toughest questions for clinicians to answer is: “How will my disease play out?” So Saria and her

colleagues decided to experiment with establishing a computational framework for predicting disease trajectories in chronic, complex diseases using EHR data. They settled on scleroderma as an interesting model disease. Scleroderma is an autoimmune disease that afflicts about 300,000 people in the United States. Some people have localized disease—hardened areas of skin in one area, perhaps; others have systemic disease. Systemic disease can progress rapidly or slowly, and it may affect the lungs, skin, gastrointestinal tract or kidneys to varying

scleroderma-related lung disease using a measure of lung health called PFVC (percent of predicted force vital capacity). Saria’s team trained a predictive model



extents. For physicians, it can be hard to know what treatments are appropriate.

Lung disease is the leading cause of death among scleroderma patients but the decline in lung function is unpredictable. So Saria’s team honed in on predicting the progression of

using data on 672 individuals collected over a period of 20 years in the Johns Hopkins Scleroderma Center patient registry. Using these data, they were able to uncover several new subtypes of lung disease progression. As time passed, the team could also dynamically personalize

predictions of lung disease progression for specific individuals. Saria has also recently shown how to account for progression in trajectories across many different organ systems in scleroderma, offering the possibility of individualizing management of systemic diseases that, like scleroderma, affect more than just one organ. Saria says the approach could be applied to other complex diseases such as asthma, autism, and cardio-obstructive pulmonary disease (COPD).

One-Button Predictions: Forecasting ALL Diagnoses

Rather than focus on individual disease risks, Page and his colleagues at the Center for Predictive Computational Phenotyping (CPCP), a Big Data to Knowledge (BD2K) Center of Excellence at the University of Wisconsin, are building a predictive model for every

current or new patients, the trained system calculates the probability each person will be assigned each diagnostic code within the next six months, Page says. The system works well even for predictions six months out, though some diseases can be predicted more accurately than others, he says.

Page hopes that the Marshfield Clinic's EHRs will start to use the system, at least for the most accurately predicted diseases. Perhaps it could offer physicians a pop-up alert if a patient crosses a threshold of risk for particular diagnoses. At the same time, he'd also like to do a careful test of whether physicians actually rely on the pop-ups. "The hope is that the prediction takes into account more features than the doctor can in one visit and can improve care," Page says.

But the work could also be useful in other ways—to help hospitals evaluate how well they are doing at treating high-risk patients system-wide, for example; or to pick potential cohorts for trials of preventive procedures; or to discover unknown long-term effects of treatments. "It could put things on the radar that aren't on there yet," Page says. "We're still at the point now where there's lots of interest and excitement about the possibilities for predictive models in the clinic, but very little translation. This work could speed up that process."

Rather than using random forests to evaluate each patient's risk of every disease, a team of researchers working with **Joel Dudley, PhD**, assistant professor of genetics and genomic Sciences at the Icahn School of Medicine at Mount Sinai in New York City, used neural networks to extract a "deep patient representation" (called Deep Patient) from 700,000 patient records in the Mount Sinai Health System's data warehouse and then tested its ability to predict the likelihood of 78 diseases in more than 70,000 patients. Shah, who did the initial data processing for the project, says Deep Patient created complex features out of the words mentioned in patient records. "It's a representation of the EHR data for risk stratification," Shah says.

Dudley's team found that Deep Patient outperformed a number of other prediction methods at predicting future assignments of disease codes. The research, which was published in *Scientific Reports* in May of 2016, could also be useful for personalizing prescriptions or recommending treatments, the paper suggests.

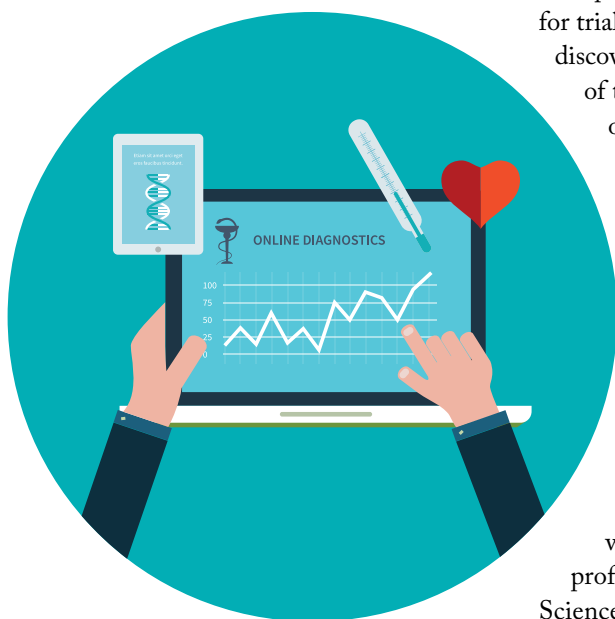
But neural nets have a downside: They don't give users an intuitive sense of what's going on. That's because they are based on finding hidden features in the data. So using Deep Patient, physicians might reliably tell patients their risk of a disease, but they wouldn't be able to point to potential reasons why.

Getting at Causation

It would be nice to go beyond predictions based on similarity to predictions based on causality, Sontag says. "The machine-learning community has, for the most part, ignored this causal inference question in recent years, but in the healthcare setting it's the most important question," he says.

Sontag and his team, including PhD student **Rahul Krishnan** and postdoc **Uri Shalit**, are currently developing several statistical approaches to discover-

ing causal relationships. One, using what's called a deep Kalman filter, is a model of disease progression that takes into consideration how drugs or treatments affect disease progression. The approach would allow researchers to ask, for example, "What would have happened to this patient if he/she had had Treatment B instead of Treatment A?" Sontag says. He's getting initial results now and says: "I view this type of work as the future of precision medicine." □



diagnosis code at a press of the button. The work relies on a high-throughput computing system called HT-Condor and the Marshfield Clinic's EHRs for more than a million patients.

To train their machine-learning algorithm, Page's team used a statistical approach called random forests—essentially a series of decision trees that identify the most informative features for each diagnostic code, then the next most informative and so on. Given a set of

