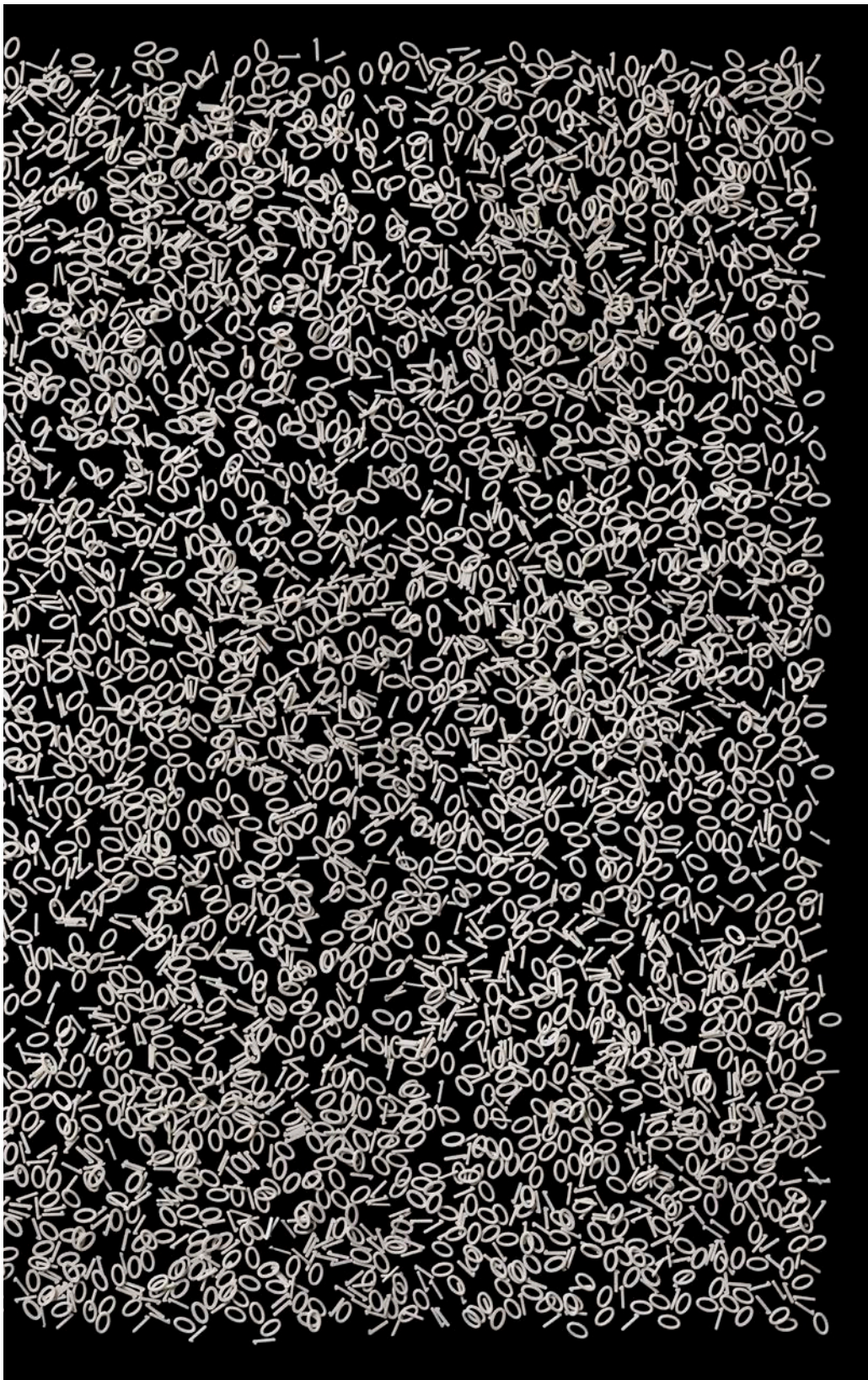
STANFORD MEDICINE

Summer 2012



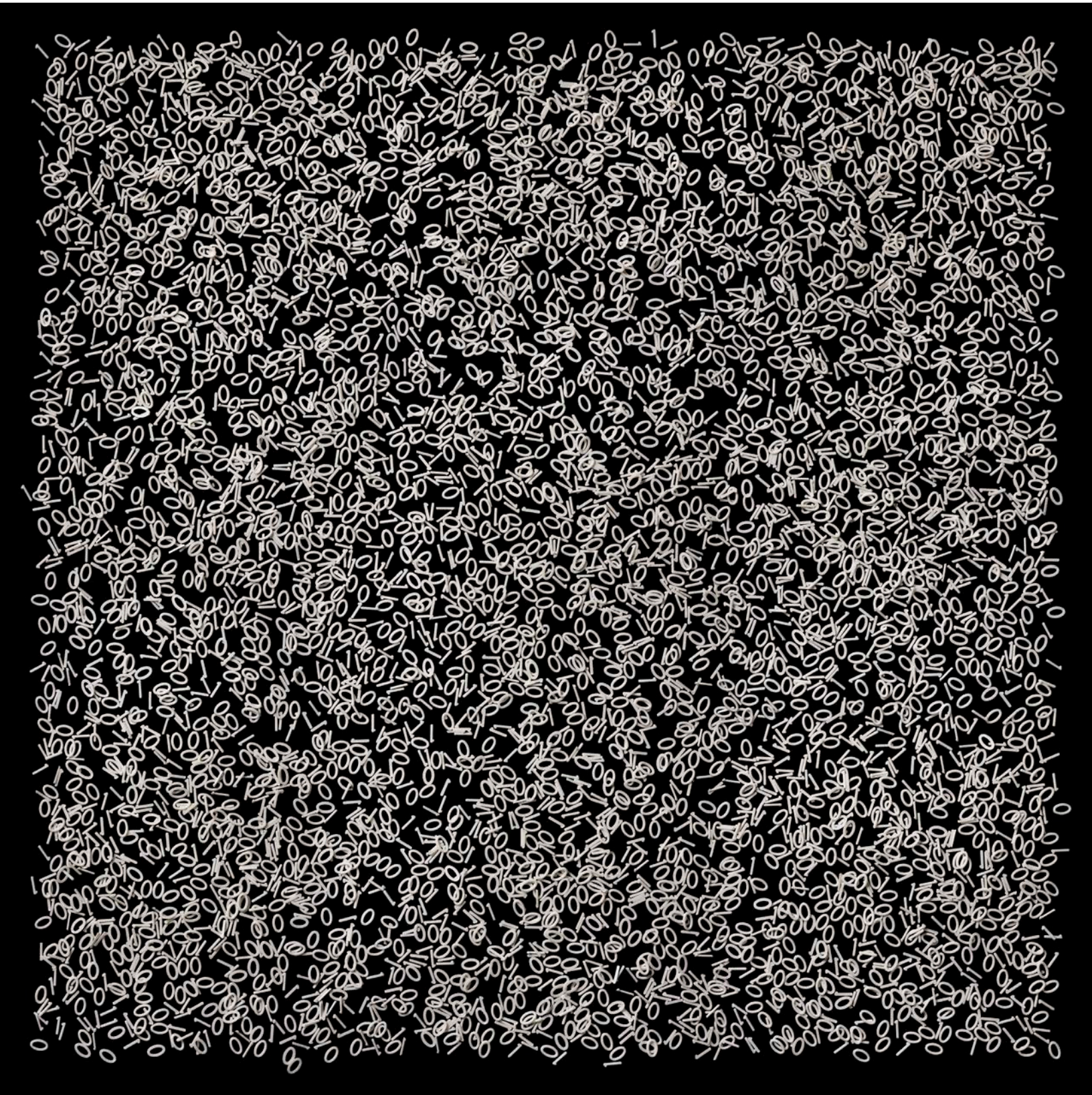
special report

DATA DELUGE

MASTERING MEDICINE'S TIDAL WAVE

STANFORD MEDICINE

Summer 2012

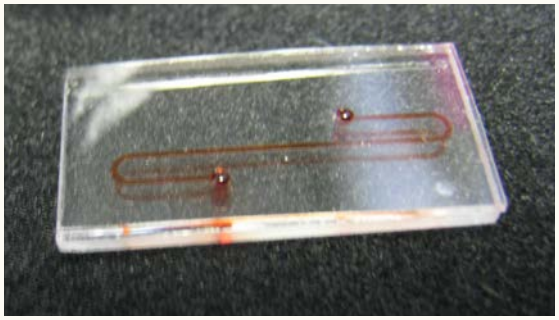


COUNT ON IT

A NEW IMMUNE-SYSTEM SENSOR COULD SPEED UP DISEASE DETECTION

Manish Butte wanted a better test for "bubble boy disease." The weeks-long wait for the standard blood test results for the severe immune deficiency was too risky, leaving newborns vulnerable to life-threatening infections. • So Butte and his team invented a new medical sensor that could detect low T-cell counts, a hallmark of the congenital immune deficiency, in just 15 minutes. That's great. But what's even better is that the sensor is so versatile and inexpensive — the prototype cost just $60 to build — it could simplify diagnosis of a huge range of diseases. • Their device, called an integrated microfluidics-waveguide sensor, sorts and counts cells in small samples of blood and other body fluids. The thumbnail-sized sensor measures different types of white blood cells, a key component of the immune system, and could be used in doctors' offices, newborn nurseries, patients' homes, disaster sites and battlefields.

"Catching infections early is important for many patient populations," says Butte, MD, PhD. He is an assistant professor of pediatrics at Stanford, a pediatric immunologist at Lucile Packard Children's Hospital and the senior author of a paper describing the sensor that appeared in March in *Biomicrofluidics*. Stanford has filed for a patent on the device; the inventors are seeking a partner to commercialize the sensors.



The sensor, about the size of an adult fingernail, offers a fast, low-cost way of simplifying the diagnosis of diseases that affect the immune system.

He hopes those who have received organ transplants, suffer chronic kidney failure or are taking immune-suppressing drugs to treat rheumatoid arthritis could use the sensors to monitor their immune systems much in the way diabetics use glucometers to track their blood sugar at home.

Each of the body's white blood cell types possesses different disease-fighting roles, and counting the cells helps doctors diagnose or monitor treatment of some diseases, including cancer and AIDS. But current cell-counting methods require fairly large blood samples and costly, slow laboratory equipment. In fact, the machines now in use cost tens of thousands of dollars.

The sensor consists of a small, rectangular piece of glass impregnated with a strip of potassium ions. The potassium-impregnated glass acts as a "waveguide" — laser light shone into the strip of glass is transmitted down it in a specific way, and the light emitted from the far end of the waveguide can be measured with a light sensor.

To operate the detector, a patient's fluid sample is mixed with antibodies specific for the particular type of white blood cell to be measured. Each antibody is attached to a tiny bead of magnetic iron. Then, the sample is injected in a small channel on top of the glass waveguide. A magnet under the glass traps the labeled cells in the channel. The iron beads block a bit of the laser light that would otherwise pass through the waveguide, and this reduced transmission is measured by the light sensor at the far side of the glass.

Among the new sensor's applications could be allowing doctors to determine the cause of a runny nose. Taking a mucus sample from the patient, doctors could use the sensor to measure the white blood cells present. Elevation of one type of cells could implicate allergies, another could point to a sinus infection and a third could suggest a common cold. — ERIN DIGITALE

PHOTOGRAPH COURTESY OF MANISH BUTTE

VOLUME 29 NUMBER 2
SUMMER 2012

STANFORD MEDICINE

The odd allure of biostatistics
page 16

SPECIAL REPORT

Data deluge

MASTERING MEDICINE'S TIDAL WAVE

Consider the yottabyte
page 6

PLUS

Take that, cancer
page 36

DEPARTMENTS

COVER: DWIGHT ESCHLIMAN. PHOTO RIGHT: COLIN CLARK

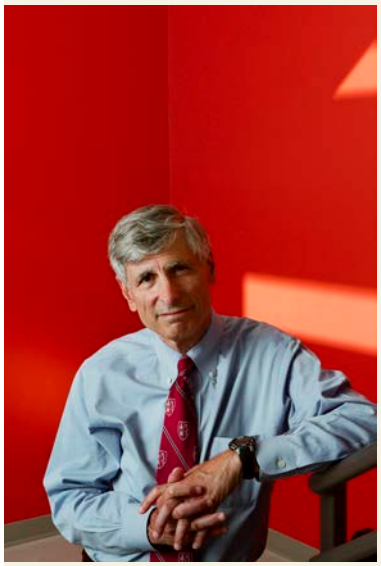
letter from the dean

It is easy to take for granted the incredible transformations in the last two decades in how we communicate, share and store information and use technology.

We have moved from high-speed Ethernet connections to wireless communications, from desktop to laptop to handheld devices, from email to texting, from personal listings to large social networks — all at remarkable speed.

The life sciences and medicine are part of this transformation. Hospitals are rapidly moving from paper to electronic medical records, and many health delivery systems give patients electronic access to their medical records and lab results. Physicians and other health-care providers can communicate with patients or review hospital records from anywhere in the world. Meanwhile, larger and larger databases are being constructed for the millions of pieces of data that comprise our genomes and other molecular profiles, and techniques are being invented to analyze and monitor these data.

These are exciting times, but they're not without danger. Amassing medical information in digital form has created the potential for security breaches with serious repercussions for individual privacy, personal security and insurability, along with enormous consequences for health-care providers and systems. The federal Health Insurance Portability and Accountability Act sets clear standards for protecting patient privacy, and it fines violators.



Yet truly safeguarding patient information has proved elusive. The Department of Health and Human Services reports 435 breaches of health privacy and security affecting more than 500 individuals since 2009 — amounting to over 20 million patients. The majority of the breaches involve electronic sources, not paper. And of the electronic source breaches, just over 60 percent involved a laptop or other portable electronic device. Like many medical institutions across the country, we have faced our own privacy breaches, and we're instituting a rigorous internal program to protect patient records.

It's worth considering what the loss of medical privacy means for an ordinary person. One outcome can be medical identity theft — the use of your name and Social Security or Medicare numbers to obtain medical care, buy drugs or submit fake billings to Medicare in your name. Aside from the disruption to your life and damage to your credit rating, such an act can be life-threatening if it leads to incorrect information appearing in your medical record. According to a study conducted last year by the Ponemon Institute, identity theft has affected about 1.5 million Americans. More fundamentally, privacy breaches further harm the trust between physician and patient.

Those of us in the medical profession need to recognize the seriousness of our responsibility. Perhaps most important is to encrypt all sensitive data stored on your laptop, smartphone or portable storage device. Never leave the device unattended (even for a moment) in a public space, especially a coffee shop, an airport bathroom or a speaker's podium. Devices left in automobiles, even in the trunk, are particularly vulnerable. And finally, unless absolutely necessary, never store sensitive information on a portable storage device in the first place.

The brave new world of data that we've entered will provide solutions to some of our most serious health issues. But like any change, benefits coexist with drawbacks. We must vigorously work to ensure that the data deluge contributes to the betterment of humankind, not its detriment.

Sincerely,
Philip A. Pizzo, MD
Dean
Stanford University School of Medicine
Carl and Elizabeth Naumann
Professor, Pediatrics, Microbiology
and Immunology

TRUJILLOPAUMIER

upfront

A QUICK LOOK AT THE LATEST DEVELOPMENTS FROM STANFORD UNIVERSITY MEDICAL CENTER

Evolution's plucky pisces

JEFFREY FISHER

TO THE UNINITIATED, the threespine stickleback might look like nothing more than a scruffy anchovy with an attitude. But this tough little fish, with its characteristic finny mohawk, is a darling of evolutionary biologists.

That's because it exhibits some of the most dramatic, adaptive changes of any animal alive today. Flourishing in fresh water or salty, appearing armored or sleek, light-skinned or dark, it's the ultimate changeling. Naturalists initially classified it as more than 50 separate species.

Now researchers at Stanford and the Broad Institute have figured out which regions of the fish's genome are responsible for the variations. The findings reverberate far beyond the stickleback: They may help scientists understand how the whale lost its hind limbs when it returned to the sea, for example, or how early humans evolved variations in skin colors.

Studying the whole-genome sequence of 21 of the fish from around the world revealed that the animals — despite their different haunts — repeatedly developed the same traits through changes in similar regions of their genomes. Specifically, the researchers identified 147 regions that varied consistently among freshwater and marine sticklebacks. About 80 percent of the changes involved regulatory regions, controlling when, where and how genes are expressed.

"This addresses a classic debate in evolutionary biology," says professor of developmental biology David Kingsley, PhD, the senior author of the study, which was published April 5 in *Nature*. "How do new traits evolve in natural populations? Do they arise through mutations in the coding regions of genes, which alter the structure and function of encoded proteins? Or are new traits the result of modifications in the regulatory regions of genes, which control where and when already-established proteins are expressed?"

In sticklebacks, at least, the balance is tilted toward the regulatory regions.

— KRISTA CONGER

Is it nuts that 100,000 medical conferences take place per year? Discuss at http://stan.md/JAcnJv.

Cancer drug redux

A TWEAK TO THE STRUCTURE of the naturally occurring cancer-curing protein interleukin-2 could overcome its major limitation: terrible side effects. It cures many cancers, but it also causes fluid to build up in the lungs and severe difficulty in breathing.

Christopher Garcia, PhD, professor of molecular and cellular physiology, and his group produced a vast variety of mutated versions of the protein, eventually obtaining one, dubbed Super-2, which in laboratory tests outperformed natural IL-2 by a wide margin in halting tumor growth. Moreover, in an animal model, Super-2 caused much less fluid buildup in the lungs. They published their study April 26 in *Nature*.

A group at the National Institutes of Health is now testing Super-2 further, in the hope of fast-tracking its development as a new therapy.

— BRUCE GOLDMAN

upfront

CAMPAIGN FOR STANFORD MEDICINE

Stanford University President John Hennessy in May announced the launch of the $1 billion Campaign for Stanford Medicine, half of which already has been committed.

The campaign will help fund the construction of a new Stanford Hospital and make investments in research and teaching. The new hospital, to be built on the current site, will replace aging facilities, incorporate advanced technologies and bring the medical center up to state seismic standards. The hospital will also greatly expand its space for treating major trauma and other emergencies.

"Providing the most advanced health care possible to people — locally, nationally and globally — will be one of the great challenges of this century," Hennessy said at the announcement. "The Campaign for Stanford Medicine draws upon our particular strengths — the proximity of the university to its hospitals and clinics — to focus on this issue and better serve the public. It will allow us to seek solutions to some of medicine's most daunting problems, and it will begin in our own community with the new Stanford Hospital."

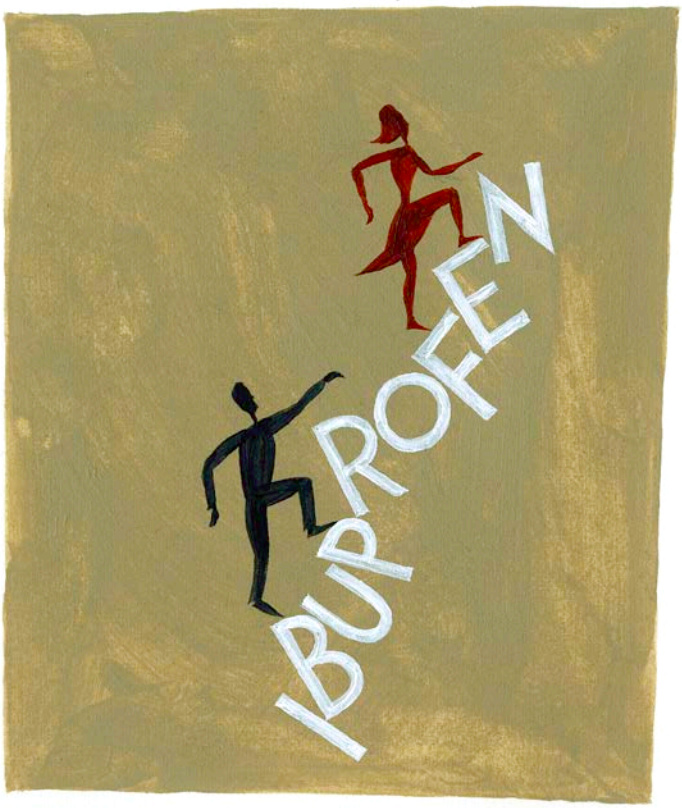
Three fundamental partners have made gifts of $50 million each to the campaign: Tashia and John Morgridge, Anne and Robert Bass, and the Redlich family. In addition, seven companies — Apple, eBay, Hewlett-Packard, Intel, Intuit, Oracle and NVIDIA — have committed a total of $175 million.

While a major portion of the campaign — $700 million — will support the new hospital, the remaining $300 million will fund research and teaching initiatives at the School of Medicine. A parallel fundraising campaign supports child health programs and an expansion of Lucile Packard Children's Hospital, which will eventually double in size.

— RUTHANN RICHTER

Altitude adjustment

"A REALLY NASTY HANGOVER" is how Grant Lipman, MD, describes the feeling of acute mountain sickness. So it shouldn't be surprising that a widely used hangover remedy, ibuprofen, prevents altitude sickness. • In a study led by Lipman, a clinical assistant professor of emergency medicine, 58 men and 28 women ascended California's White Mountains to test ibuprofen's impact on the condition, which in severe cases can be fatal if untreated. They started at 4,100 feet, taking either 600 milligrams of ibuprofen or a placebo at 8 a.m., and headed up in cars to 11,700 feet, where they had a second dose at 2 p.m. Then they hiked to 12,570 feet and had a third dose at 8 p.m., before spending the night on the mountain. • Ibuprofen cut the incidence of altitude sickness by 26 percent and reduced symptoms overall. The study was published in the June issue of *Annals of Emergency Medicine*. — JOHN SANFORD



Stop lecturing me

Despite dramatic changes in the medical world over the past century, medical education still relies on the traditional lecture format. It's time for that to change, say the authors of "Lecture halls without lectures," in the *New England Journal of Medicine*.

In the May article, co-author Charles Prober, MD, senior associate dean for medical education, proposes replacing most lectures with short online videos, which would free up class time for more interactive education.

"Teachers would be able to actually teach, rather than merely make speeches," the authors write.

Stanford's core biochemistry class tested the method this year: Students watched short videos on their own time and used classes to discuss clinical vignettes highlighting the biochemical bases of diseases.

"Student reviews of the course improved substantially from the previous year, and class attendance increased from 30 to 80 percent, even though class attendance was optional," they write.

— TRACIE WHITE

ILLUSTRATIONS: JEFFREY FISHER; PHOTOGRAPH: STEVE FISCH

LOW-COST CANCER SCAN

TENS OF THOUSANDS OF PEOPLE DIE NEEDLESSLY EACH YEAR IN DEVELOPING COUNTRIES FROM MOUTH CANCERS that could have been detected early with regular dental checkups. But with fewer than one dentist per 100,000 people in the world's rural areas, that's not an option. Now an ultra-low-cost device being developed at Stanford and tested this summer in India may enable early diagnosis.

Invented by Manu Prakash, PhD, assistant professor of bioengineering, and his team, the device attaches to a smartphone's built-in camera, providing a high-resolution, panoramic image of the inside of a mouth. Illuminated by a blue fluorescent light, malignant cancer lesions show up as dark spots on the image, which could be sent to dentists or oral surgeons wirelessly.

The scanner is designed for mass production, with an estimated material cost of just a few dollars.

— KRIS NEWBY

Zombies among us

NEW STANFORD RESEARCH shows that about 3.6 percent of American adults — or around 8.4 million — are prone to sleepwalking. The study is the first to use a large, representative sample of the U.S. general population to demonstrate the number of sleepwalkers (19,136 adults in 15 states were surveyed), and the researchers say the findings underscore the fact that sleepwalking is much more prevalent in adults than previously appreciated. • Maurice Ohayon, MD, DSc, PhD, a professor of psychiatry and behavioral sciences, and his colleagues also found an association between nocturnal wanderings and certain psychiatric disorders, such as depression and anxiety. They published their study May 15 in the journal *Neurology*.

— MICHELLE L. BRANDT

29% of adults surveyed reported sleepwalking at least once in their lives.

Heart-test excess

Left ventriculography, an invasive imaging procedure that measures heart function, is dramatically overused, according to a new Stanford study. The procedure adds both risk and cost, says co-author Ronald Witteles, MD, assistant professor of cardiovascular medicine, calling the rates of unnecessary use of the test "shockingly high."

The procedure was developed 50 years ago to assess heart function by measuring its ejection fraction — the percentage of blood in the heart's left ventricle that gets squeezed out with each beat. But over the years, several less-invasive and often superior methods of measuring ejection fraction have emerged, such as echocardiograms and nuclear cardiac imaging.

For the new study, the researchers examined a national database of about 96,000 patients who in 2007 had a coronary angiography — a procedure for imaging the coronary arteries. During a coronary angiography, cardiologists insert a catheter into a blood vessel in the patient's arm, groin or neck, thread it into a coronary artery, and fill it with dye to release in the bloodstream. Cardiologists perform left ventriculography, which costs about $300, as an add-on to angiography, advancing a catheter from the artery into the left ventricle.

Suspecting that left ventriculography was overused, the researchers zeroed in on the 37,000 coronary angiography patients who had previously undergone one of the new methods of measuring ejection fraction. Surprisingly, 88 percent of those patients went on to get a left ventriculography.

"If a patient recently had an echocardiogram or a nuclear study, it didn't make them less likely to have the left ventriculography procedure — it made them more likely," Witteles says. "That is impossible to explain from a medical justification standpoint."

The study was published in the *American Heart Journal* in April.

— TRACIE WHITE

B ! G D A T A

WHAT IT MEANS FOR OUR HEALTH AND THE FUTURE OF MEDICAL RESEARCH

By Krista Conger

PHOTO-ILLUSTRATION BY DWIGHT ESCHLIMAN

PHOTOGRAPHY BY COLIN CLARK

"I think I'm getting sick," says Michael Snyder, PhD, cheerfully, unbuttoning his shirt cuff and rolling up his sleeve to show me the inside of his elbow. • A ball of white cotton puffs out from under the white tape used to cover the site of a blood draw. Snyder, 57, a lean, intense-looking man who chairs Stanford's genetics department, doesn't sound at all upset about his impending illness. In fact, he sounds positively satisfied. "It's a viral illness; one of the hazards of having young children, I guess," he says, smiling. "So far I'm only a little sniffly. But it does give me another sample." • I sat in a chair in his small office, pondering the handshakes we had exchanged moments earlier and surreptitiously wiping my palm on my jeans. • Blood samples are nothing new to any of us. They're a routine part of medical checkups and give vital information to clinicians about our cholesterol and blood sugar levels, the function of our immune systems, and our production of hormones and other metabolites.

The fleeting pinch of discomfort is a small price to pay for such a peek inside our own bodies.

But Snyder's recently obtained blood sample was destined for a different fate. It was also what had brought me to his office on an unseasonably warm afternoon in February. While the standard blood test panel recommended during routine check-ups assesses the presence and levels of only a few variables, such as the total numbers of red and white blood cells, hemoglobin and cholesterol levels, Snyder's would undergo a much more intense analysis — yielding millions of bits of data for an ultra-high-definition portrait of health and disease.

The unprecedented study, termed an integrative personal genomics profile, or iPOP, generated billions of individual data points about Snyder's health, to the tune of about 30 terabytes (that's about 30,000 gigabytes, or enough CD-quality audio to play non-stop for seven years).

It's a lot of data. And he's just one man, with just a few month's worth of samples. Other biological databases that are frequently used, and still growing, include an effort to categorize human genetic variation (more than 200 terabytes as of March) and another to sequence genomes of 20 different types of human cancer (300 terabytes and counting).

Which begs the question: What wonderful things can we do with all this data? How do we store it, analyze it and share it, while also keeping the identities of those who provided the samples private and protecting them from discrimination? Meeting this challenge may be the most pressing issue in biomedicine today.

WE'RE DROWNING IN DATA. **Supermarkets, credit cards, Amazon and Facebook. Electronic medical records, digital television, cell phones.** The universe has gone wild with the chirps, clicks, whirs and hums of feral information. And it truly is feral: According to a 2008 white paper from the market research firm International Data Corp., the amount of data generated surpassed our ability to store it back in 2008. The cat is out of the bag.

But what are we talking about, really? To truly understand the issue, it's necessary to do a little background work. Computers store data in a binary fashion with bits and bytes. A bit is defined by one of two possible, mutually exclusive states of existence: up or down, for example, or on or off. In computers, the states are represented by 0 and 1. A byte is the number of bits (usually eight) necessary to convey a unit of information, such as a letter of text or a small number on which to perform a calculation. A kilobyte is 1,000 bytes; a megabyte is 1,000 kilobytes.

The numbers pile up relentlessly: giga, tera, peta, exa, zetta, which is 1 sextillion bytes, up to yotta, which is officially too big to imagine, according to *The Economist*. If you want to try to imagine it anyway, a yottabyte equals about a septillion bytes. You write septillion as a 1 followed by 24 zeroes.

In 2010, the amount of digital information — from high-definition television signals to Internet browsing information to credit card purchases and more — created and shared exceeded 1 zettabyte for the first time. In 2011, it approached 2. The amount has grown by a factor of nine in five years, according to IDC, which pointed out in its 2011 report that there are "nearly as many bits of information in the digital universe as stars in our physical universe."

In March, the federal government announced the Big Data Research and Development Initiative — a $200 million commitment to "greatly improve the tools and techniques needed to access, organize and glean discoveries from huge volumes of digital data." In particular, the National Institutes of Health and the National Science Foundation are offering up to $25 million for devising ways to visualize and extract biological and medical information from large and diverse data sets like Snyder's study, and simultaneously the NIH announced it would provide researchers free access to all 200 terabytes of the 1,000 Genomes Project — an attempt to catalog human genetic variation — via Amazon Web Services.

Successfully managing the "data deluge" will allow scientists to compare the genomes of similar types of cancers to identify how critical regulatory pathways go awry, to ferret out previously unknown and unsuspected drug interactions and side effects, to precisely track the genetic changes that have allowed evolving humans to populate the globe, and even to determine how our genes and environment interact to cause obesity, osteoporosis and other chronic diseases.

MIKE SNYDER'S "OMICS STUDY" IS STILL ONLY A TINY FRACTION OF THIS TOTAL. **But the iPOP represents on a small scale the complex challenges of accumulating, storing, sharing and protecting biological and medical data —** as well as the need to extract useful clinical information on an ongoing basis. And we'd better figure out what we're doing.

"This is going to be standard medical care," says Snyder. "It will change the way we practice medicine." If he is right, we're looking at the ongoing accumulation of terabytes of health information for each of us during our lifetimes.

That's because the study's outcome (published in *Cell* in March) is like nothing medicine has ever seen before. It's a

MIKE SNYDER, CHAIR OF GENETICS,
has amassed billions of data points in his health profile.


NO FOOD OR DRINKS
NALGENE

series of snapshots that show how our bodies use the DNA blueprints in our genomes to churn out RNA and protein molecules in varying amounts and types precisely calibrated to respond to the changing conditions in which we live.

The result is an exquisitely crafted machine that turns on a dime to metabolize food, flex muscles, breathe air, fight off infections and make all the other adjustments that keep us healthy. A misstep can lead to disease or illness; understanding this dance could help predict problems before they start.

SO WHERE'S ALL THIS DATA COMING FROM, AND WHERE DO WE KEEP IT? **Great leaps in scientific knowledge have almost always been preceded by improvements in methods of measurement or analysis.** The invention of the optical microscope opened a new frontier for biologists of the early to mid-1600s; the telescope performed a similar feat for astronomers of the period. Biologists of the 20th century have their own touchstones, including several developed at Stanford: the invention of the fluorescence-activated cell sorter around 1970 by Leonard Herzenberg; the invention of microarray technology in Patrick Brown's and Ronald Davis's labs in the mid-1990s, and the recent development of ultra-high-throughput DNA sequencing technology and microfluidic circuits by bioengineer Stephen Quake. Each of these technological leaps sparked rapid accumulation of previously unheard of volumes of data — and a need for some place to store it.

One of the first publicly available databases, Genbank, started in 1982 as a way for individual researchers to share DNA sequences of interest. In November 2010, the database, which is maintained by the National Center for Biotechnology Information, contained hundreds of billions of nucleotides from about 380,000 species. It has doubled in size approximately every 18 months from its inception until 2007.

Other databases have sprung up over the years, including OMIM, or Online Mendelian Inheritance in Man, which catalogs the relationship between more than 12,000 genes and all known inherited diseaseand GEO, or Gene Expression Omnibus, which contains over 700,000 RNA sequences and microarray data that illuminate how the instructions in DNA are put into action in various tissues and at specific developmental times.

But we can't stop there. It's also necessary to know how gene expression is affected by outside forces like diet or environment, and how the protein product of that gene interacts with the hundreds or thousands of other proteins in the cells to achieve a certain biological outcome. And then there's the issue of time, which, according to Douglas Brutlag, PhD, a bioinformatics expert and emeritus professor of biochemistry and medicine at Stanford, is "the worst dimension of all."

"Ideally, we would like to have a database of all human protein-coding genes, of which there are about 22,000," says Brutlag. "We'd include information about their levels of expression in every tissue (about 200) and in every cell type (about 10 to 20 in each tissue). And then we'd need to know exactly when these genes were expressed in each of these locations during development. You start multiplying these variables together, and then you see, 'Now that's big data.'" It can also be a big headache.

WHEN HE FIRST PROFFERED HIS ARM IN JANUARY 2010, **Mike Snyder wasn't thinking of the volume of data his iPOP study might generate.** He was focused mostly on the type of information he and his lab members would receive.

"Currently, we routinely measure fewer than 20 variables in a standard laboratory blood test," says Snyder, who is also the Stanford W. Ascherman, MD, FACS, Professor in Genetics. "We could, and should, be measuring many, many thousands. We could get a much clearer resolution of what's going on with our health at any one point in time." But the process is constrained by habit and infrastructure — standard medical labs are not set up to do anything like Snyder's iPOP, and many primary care physicians would have no idea what to do with the information once they had it. It's also expensive: It costs about $2,500 to extract all the bits of information from each of Snyder's blood samples.

'WE'VE BEEN SO FOCUSED ON GENERATING HYPOTHESES, BUT THE AVAILABILITY OF BIG DATA SETS ALLOWS THE DATA TO SPEAK TO YOU. MEANINGFUL THINGS CAN POP OUT THAT YOU HADN'T EXPECTED.'

The first step of Snyder's iPOP was relatively simple: He had his whole genome sequenced. What was a daunting task fewer than two decades ago is now a straightforward, rapid procedure with a cost approaching $1,000. Furthermore, the whole-genome sequence of any one individual doesn't actually take up that much room in our universe of digital storage. The final reference sequence compiled by the Human Genome Project can be stored in a modest 3 gigabytes or so — a small portion of a single hard drive.

But while pulling out differences between a sequence from an individual like Snyder and the reference sequence is simple, ascertaining what those differences mean is another matter entirely. The layers of annotation, describing what genes

are involved in which processes, add complexity. For that, it's necessary to understand which variation is normal, which is inconsequential and which might be associated with disease. Here, the field of bioinformatics can help.

Bioinformaticians marry computer science with information technology to develop new ways to analyze biological and health data. The field is one of the most rapidly growing in medicine and includes Stanford researchers Atul Butte, MD, PhD, and Russ Altman, MD, PhD, each of whom has developed computerized algorithms to analyze publicly available health information in large databases. These algorithms enable researchers to approach the data without preconceived ideas about what they might find.

"We've been so focused on generating hypotheses," says cardiologist Euan Ashley, MD, "but the availability of big data sets allows the data to speak to you. Meaningful things can pop out that you hadn't expected. In contrast, with a hypothesis, you're never going to be truly surprised at your result." Hypothesis-driven science isn't dead, say many scientists, but it's not the most useful way to analyze big data sets.

Ashley and colleagues, including Snyder, Altman and Butte, have designed a RiskOGram computer algorithm designed to incorporate information from multiple disease-associated gene variants in an individual's whole-genome sequence to come up with an overall risk profile. They first used the algorithm to analyze the genome of Quake, their bioengineering colleague. Since then it's been used on whole-genome sequences from several additional people, including the family of entrepreneur John West and, most recently, Snyder himself. Together, Altman, Ashley, Butte, West and Snyder have launched a company called Personalis to apply genome interpretation to clinical medicine.

When thinking about the storage of sequencing data, however, there's also the issue of sequencing "depth," which essentially means the number of times each DNA fragment is sequenced during the procedure. Repeated sequencing reduces the possibility of error. Extremely deep sequencing, like that employed with Snyder's genome (each nucleotide was sequenced, or "covered," an average of 270 times) comes at a price beyond dollars; it generates a plethora of raw data for storage and analysis. In comparison, the Human Genome Project provided about eightfold coverage of each nucleotide in the reference sequence.

Finally, today's sequencing technologies don't sequence an entire chromosome in one swoop, but first break it into millions of short, random fragments. This leads to another type of biological data to store. After each fragment is sequenced, computer algorithms assemble the small chunks into chromosome-length pieces based on bits of sequence overlap among the fragments. It's necessary to save the raw data to closely investigate any discrepancies in critical disease-associated regions.

What's it all add up to? Well, a "normal" genome might require only a few gigabytes to store. But Snyder's, with the deep coverage, raw data and extensive annotation, takes up a whopping 2 terabytes, or about 2,000 gigabytes — plenty of fodder for the RiskOGram.

"I WAS AMUSED TO SEE TYPE-2 DIABETES EMERGING SO STRONGLY," **says Snyder of the results of the analysis, who, despite his fondness for cheeseburgers and mint chocolate chip ice cream had no known risk factors or family history of the disease.** In contrast, he was lean and relatively active. Despite this, the RiskOGram predicted that his risk of developing type-2 diabetes was 47 percent, which is more than double that of other men his age.

Snyder also learned he had a genetic predisposition to basal cell carcinoma and a propensity for heart disease (a finding that caused him to start cholesterol-lowering medication). That last finding wasn't unexpected; many family members on his father's side had died from heart failure.

What Snyder was doing was, in principle, not unusual. It's just very thorough. Companies like 23andMe allow consumers to send in a cheek swab and receive a report of disease risk based on regions of genetic variation called single nucleotide polymorphisms, or SNPs. Unlike whole-genome sequencing, tests of this type scan the genome for the presence of variations already known to affect health or disease risk.

As a precaution, Snyder decided to include regular blood sugar level tests as part of his iPOP — high glucose indicates diabetes. He also scheduled a meeting with Stanford endocrinologists and diabetes experts Sun Kim, MD, and Gerald Reaven, MD, who suggested a rigorous three-hour fasting blood sugar test. Snyder agreed, but his hectic schedule delayed the test by about five months. A more pressing issue was how to upload and store the information generated by his iPOP to the appropriate public databases. The sheer volume of the data complicated the task enormously.

"The hard part was submitting all the data," says Snyder. "Transferring these large files takes time. The whole process of submission took several weeks." Since the publication of Snyder's iPOP study, many other researchers have requested access to the raw data. In most cases, Snyder says, it's been far easier to simply ship a hard drive containing the data to the requesters rather than transmit them electronically.

SINCE THE INCEPTION OF GENBANK AND OMIM, **federal biological databases have gotten progressively more ambitious.** The 1,000 Genomes Project, for example, aims to curate and catalog naturally occurring

differences in the genomes of 2,600 ethnically diverse humans (as of March, the project had collected more than 200 terabytes of data on about 1,700 humans, all publicly accessible).

And then there is the Cancer Genome Atlas, or TCGA — an effort to sequence the entire, error-riddled genome of tumor cells from 20 human cancers. Scientists involved in the effort, funded by the National Cancer Institute and the National Human Genome Research Institute, have sequenced the genes from more than 1,000 tumor genomes. Furthermore, researchers typically collect two genomes from each patient: one from normal tissue and one from the tumor. The TCGA project does very deep sequencing of the genes in both samples, coupled with complete genome sequencing for many samples. Total data produced as of May 2012 is more than 300 terabytes.

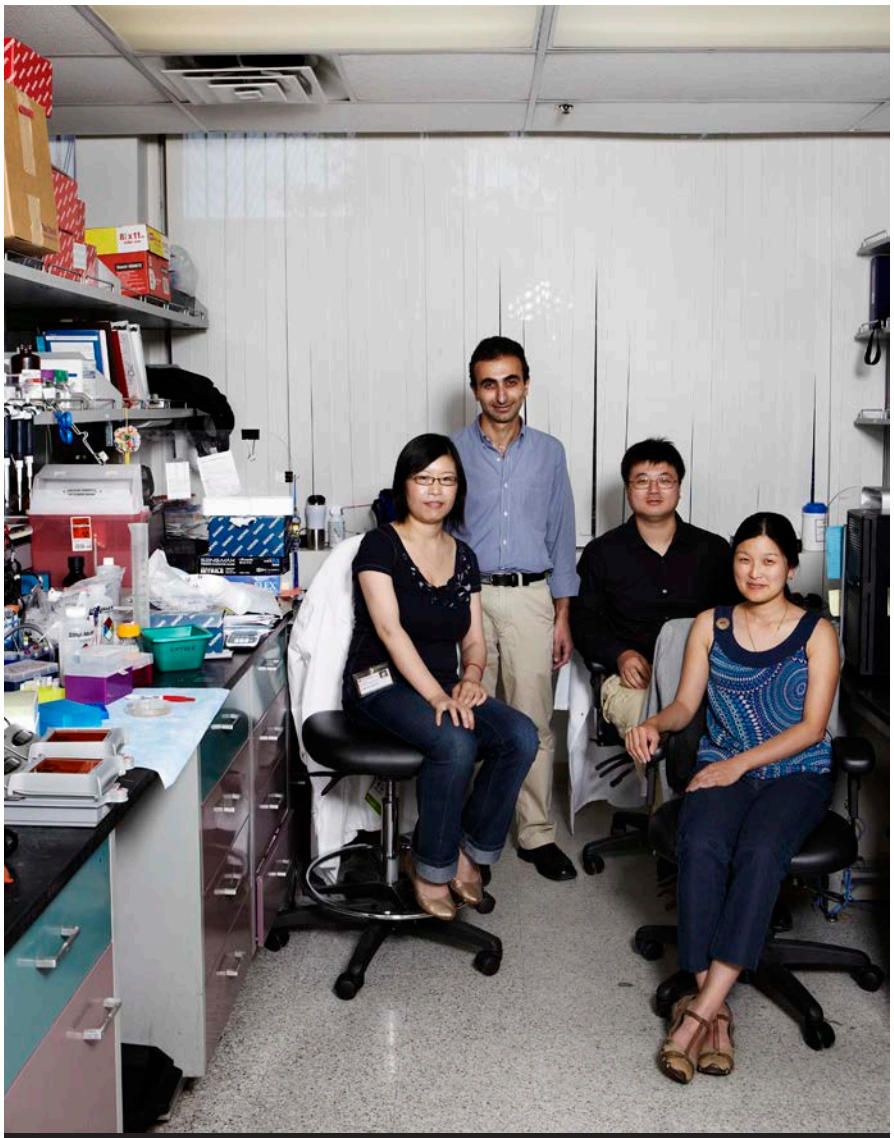
"Just moving these files from one institution to another is a major headache," says David Haussler, PhD, a distinguished professor of biomolecular engineering at the University of California-Santa Cruz and a co-principal investigator of one of TCGA's seven Genome Data Analysis Centers. He and his colleagues are developing new algorithms and software to transfer and analyze the resulting data.

In May, Haussler's Santa Cruz group announced the completion of the first step in the construction of the Cancer Genomics Hub — a large-scale data repository and user portal for the National Cancer Institute's cancer genome research programs. Based in San Diego, the hub will support not just TCGA, but also two related large-scale projects: one (called TARGET) focused on five childhood cancers and another (called CGCI) focused primarily on HIV-associated cancers.

"In addition to being one of the genome data analysis centers for TCGA, our Cancer Genomics Hub will store the sequences themselves," says Haussler. "All of these data will be gathered into one place and made available to researchers."

Comparing the genomes of thousands of cancers will allow researchers to distinguish meaningful differences in the hundreds of mutations harbored by each tumor and to develop therapies that target these mutations. It's critical to cancer research because the immense variation among and even within individual cancers stymies traditional research conducted on panels of tumors handily available at a scientist's institution. Instead it's necessary to compare thousands or tens of thousands of tumors to identify mutations that drive the disease and its response to therapy.

Haussler and his collaborators designed the San Diego facility to handle 5 petabytes of information and to be scalable to handle 20. "In the last few years the TCGA project has produced more data than the combined total of all other DNA sequencing efforts accumulated by NCBI from its inception as Genbank in 1982 until January of 2012," says Haussler. "TCGA will generate orders of magnitude more data beyond this. If we had the money, we'd certainly be sequencing multiple biopsies, or, perhaps one day, even



L TO R: LIHUA JIANG, GEORGE MIAS, RUI CHEN AND JENNIFER LI-POOK-THAN co-authored the paper describing the analysis of Mike Snyder's health data.

individual cells from each tumor."

According to Haussler, the amount of information generated by the researchers working on the TCGA project is comparable to that generated by the Large Hadron Collider particle accelerator. "In cases like this, it's much simpler to bring the computation to the data, than the other way around," says Haussler. In many cases he and his colleagues suggest that researchers interested in conducting repeated studies co-locate some of their computing equipment in the San Diego Supercomputer Center alongside the Cancer Genomics Hub — a riff on Snyder's smaller-scale solution of mailing hard drives to researchers. "Researchers bringing computation to the data either run analysis on their own custom computing platforms located near the data or 'rent' time on generic facilities — the cloud computing model," says Haussler. Cloud computing — in this case, hosting the computation on server farms that have access to the data — could be the answer for some databases. Security and privacy concerns could be a roadblock to cloud computing for some uses, especially when patient information is involved.

"Yes, there are a lot of obstacles," says Haussler. "Chief among them are the sheer size of the data involved, the issues of privacy and compliance, and the natural reluctance of some researchers to share their data with one another. But dammit, we're overcoming them. It's getting better all time."

"THERE'S NO WAY YOU HAVE DIABETES," **Sun Kim told Snyder in February 2011 as the two walked back to a phlebotomy chair at Stanford Hospital & Clinics.** Snyder's packed schedule had finally allowed him to squeeze in the three-hour fasting blood sugar test recommended by Reaven. Snyder agreed. He'd been performing quick tests of his own levels for months by that time.

"Sometimes I fasted, sometimes I didn't," says Snyder of the tests. "It didn't seem to matter. My glucose levels were always very normal, even when I'd just eaten a cheeseburger and fries for lunch. There was never any indication of a problem."

And his overall health had remained good, with the exception of a recent viral illness he probably contracted from one of his two daughters, Emma (now age 11) and Eve (age 6). (He'd missed several days of work as a result.) But he pressed forward in the name of science.

"There are some indications in my genome that I may be at risk," he reminded Kim, while settling into the chair. "Let's go ahead with the test."

And then the real surprises started.

Snyder's initial blood glucose level was 127. Normal fasting glucose levels range between 70 and 99 milligrams per deciliter.

"We were both really surprised," says Snyder. "I didn't have my data sheets with me at the time, but I knew it had always been quite low. So Kim repeated the analysis." The result was the same. The three-hour test requires the patient to drink a sugary drink and give three subsequent readings as the body metabolizes the sudden influx of sugar. The results showed that, although Snyder's initial blood sugar levels were high, he was still able to metabolize the sugar normally. But that jarring initial reading lingered in Snyder's mind.

AT THAT TIME, **Snyder's iPOP, or "personal 'omics' profile," study had been going for about nine months.** The word "omics" indicates the study of a complete body of information, such as the genome (which is all DNA in a cell), or the proteome (which is all the proteins). Snyder's iPOP also ncluded his metabolome (metabolites), his transcriptome (RNA transcripts) and autoantibody profiles, among other things. It's a dynamic, ongoing analysis that accumulates billions of additional data points with each blood sample he contributes. He was contributing a sample for analysis about every two months, more often when he became ill.

With each sample Snyder gave (about 20 in all over the course of the initial study, and 42 samples as of early June 2012), the researchers took dozens of molecular snapshots, using different techniques, of thousands of variables and then compared them over time. The composite result was a dynamic picture of how his body responded to illness and disease.

Some of these snapshots included sequencing the RNA transcripts that carry the instructions encoded by DNA in a cell's nucleus out to the cytoplasm to be translated into proteins. Par for the course, the researchers completed this task at an unheard-of level of detail, sequencing 2.67 billion snippets of RNA over the course of the study. These sequences allowed them to see what proteins or regulatory molecules Snyder's body was producing at each time point and, similar to looking at a list of materials at a construction site, to infer what was going on inside his cells. Overall, the researchers tracked nearly 20,000 distinct transcripts coding for 12,000 genes — each submitted to the GEO database — and measured the relative levels of more than 4,000 proteins and 1,000 metabolites in Snyder's blood. In particular, they measured the levels of several immune molecules called cytokines, which changed dramatically during periods of illness like the one he'd experienced two weeks before his fasting glucose test.

Snyder believed in submitting his data to public databases because it stands to help other investigators. In many cases it's also required prior to scientific publication of any subsequent results. But private databases have proven they can also facilitate research. According a press release from 23andMe in June 2011, the company had approximately 100,000 customers, of which more than 76,000 had consented to have their de-iden-

tified information used for scientific research. For example, scientists compared data from over 3,000 patients with Parkinson's disease with those of nearly 30,000 existing 23andMe clients without the disease. They found two novel, previously unsuspected gene variants associated with the disorder.

"The fact that we are searching these genomes for recurring patterns without any preconceived notions about what we might find is what's so fundamentally important about this approach," says Haussler. "In this way, we are able to recognize any gene that's recurrently mutated in cancer, for example, regardless of any expectations we may have."

The approach works well for other databases too. Stanford's Altman, MD, PhD, a professor of bioengineering, of genetics and of medicine, took a similar tack to reveal unexpected drug side effects and drug-drug interactions by using the Food and Drug Administration's Adverse Event Reporting System database and a Stanford database of patient information called STRIDE. *[See story, page 28.]* The Veterans Affairs Administration's VistA electronic health information system, which contains de-identified information on millions of patients, is also a valuable source of data for researchers.

Big data of all sorts — particularly of the type generated by Snyder's iPOP — is important to researchers seeking to understand the complex, interconnected dance of the thousands of molecules in each cell. It's a new type of research known as systems biology.

"We got stuck for a while in the idea of direct causality," says Ashley, the cardiologist. "We thought we could turn one knob and measure one thing. Now we're trying to really understand the system. We're developing an appreciation for the 10,000 things that happen on the right when you poke the network on the left." Ashley estimates that there are about 4.3 X 10^67, or 43 "unvigintillion" (that's 43 followed by 66 zeros) possible 20-member gene networks — that is, collections of pathways or actions that accomplish similar goals — within a cell.

Seattle-based nonprofit SAGE Bionetworks, started by Stephen Friend, MD, PhD, and bioinformatics whiz Eric Schadt, PhD, aims to help individual researchers tackle systems biology by pooling their data and developing shared computer algorithms to analyze it. It's created Synapse, an online service that the company bills as an innovation space, providing researchers with a web portal, data analysis tools and scientific communities to encourage collaboration.

SNYDER STARTED THE IPOP AS A WAY **to apply the complex concepts of systems biology to clinical outcomes.** But at the moment, he was primarily concerned with just one number: his hemoglobin A1C, which is a more sensitive analysis of glucose levels over a period of weeks. One week after his fasting blood sugar test, he found that his hemoglobin A1C level was also elevated: 6.4 percent (normal ranges are between 4 and 6 percent for non-diabetic people). "At this point, my wife was starting to get a little concerned," says Snyder.

As was he. Snyder called his mother, a retired schoolteacher in Pennsylvania, to ask a few more questions. She recalled that his grandfather had been diagnosed with high blood sugar levels late in his life. But that seemed minor. "He cut out desserts, and lived to be 87 years old," says Snyder. All in all, nothing that would automatically raise red flags. But there was that initial fasting glucose level. And the elevated A1C.

So he lined up an appointment with his regular physician about six weeks later. "She, too, took one look at me when I walked in and said, 'There's no way you have diabetes,'" says Snyder. But his hemoglobin A1C level at that appointment was 6.7, crossing the threshold for diagnosis. He was diabetic.

"I remember getting the confirmation with the second A1C test," says Snyder. "It left me very conflicted. Scientifically, it was quite interesting, but personally it's not news anyone wants to hear. I had a serious health issue for the first time in my life."

THERE WAS NEVER ANY QUESTION **that Snyder would submit his personal health information to public databases like Genbank and GEO.** Of course he would. And he wasn't particularly worried about privacy. He has tenure, and relatively good health insurance benefits. But when he visited his primary care physician, he crossed a line. Now his diabetes diagnosis and its implications were medically codified.

"My wife is 10 years younger than I am," says Snyder, "and we have two young daughters." As a result, she began looking into increasing his life insurance policy.

"The life insurance company quoted her an additional $7,000 fee because of my physician's diabetes diagnosis," says Snyder. "When my wife asked what would happen if I were able to get my glucose levels down, they told her it didn't matter. Once you're diabetic, you're always diabetic, in their eyes. Lower glucose levels just means you've managed your disease."

Even people with identified genetic risk but no diagnosis may run into trouble: Although the 2008 passage of the Genetic Information Nondiscrimination Act prohibits health insurance companies and employers from discriminating on the basis of genetic information, no such protection exists when it comes to life insurance or long-term disability.

The issue illustrates the delicate balance that must be struck when using biological and health information in clinical decision making and research and shows why many people may be leery about consenting to share their private data. Conversely, researchers working with big data can often find themselves bushwhacking through what seems like an impenetrable thicket of consent forms and privacy concerns.

"Working through issues of compliance with the National Institutes of Health for the data in the TCGA Cancer Genome Hub was like nothing I've ever dealt with before," says Haussler. "Until we come up with an acceptable, more flexible compliance infrastructure, it's going to be very challenging to move forward into cloud computing." Issues include how to ascertain who should be able to access the data, and how anonymity can be preserved in the face of genomic sequencing that, in its very nature, identifies each participant.

But things may be about to change. Several groups, including the nonprofit SAGE, are working on a standardized patient consent form. This will allow people contributing personal health information to shared databases to give a single consent to a plethora of research activities conducted on their pooled, anonymous data. Sometimes also called portable legal consent, the effort is essential to streamline research while also protecting patient privacy and prohibiting discrimination. Another effort, MyDataCan, based at Harvard, aims to provide a free repository for individual health data. Using the service, a participant can pick and choose which research activities to grant access to their personal data.

Conversely, pending legislation in the California Senate called the Genetic Information Privacy Act would squelch research by requiring consent by individuals for every research project conducted on their DNA, genetic testing results and even family history data, according to *Nature* magazine. Without such authorization, researchers would have to destroy DNA samples and data after each study — making it impossible for scientists to use genetic databases for anything other than the original purpose for which the information was gathered.

Although it's clear that issues of privacy still need to be addressed, it's equally clear that public access to this and other types of biological data can be very useful. And it's not just researchers who can contribute.

Crowdsourcing is a buzzword that's become reality in some fields of biology. One example of crowdsourcing, FoldIt, is a game developed by the Center for Game Science at the University of Washington that allows anyone with a computer to solve puzzles of protein folding. Tens of thousands of nonscientists around the world have rejiggered protein chains into a nearly infinite number of possible structures.

Stanford-designed EteRNA is another example: Players design and fold their own RNA molecules. And Atul Butte mentors high school and college students in his lab as they use information from publicly available tumor databases to identify important genes and pathways. *[See story, page 20.]*

It's a new era — one in which the public and scientists collaborate to plumb vast amounts of data for biological and medical insights. The authors of scientific papers using big data increasingly number in the tens or even hundreds as researchers around the world pool their data and work together to find meaningful outcomes. The issue of who stands to benefit commercially from the results of such research is also unresolved — in May, 23andMe filed a patent related to its discovery of one of the two gene variants associated with Parkinson's disease, to the consternation of some research participants and scientists who feel that the company is looking to benefit financially from data freely shared by participants. The company, in turn, argues that such patents are necessary to encourage drug development that leads to new therapies for patients. But, regardless of legal tussles, the results of such analyses stand to affect individuals like you and me in a way never before imagined.

ONE WEEK AFTER HIS FASTING BLOOD SUGAR TEST, HE FOUND THAT HIS HEMOGLOBIN A1C LEVEL WAS ALSO ELEVATED. 'AT THIS POINT, MY WIFE WAS STARTING TO GET A LITTLE CONCERNED.'

MIKE SNYDER DOESN'T EAT ICE CREAM ANYMORE. **Gone are his trips to snag candy bars from hallway vending machines and the lunches of burgers and fries.** He went cold turkey on April 13, 2011. He rides his bike to work nearly every day, and has taken up running again. He lost 15 pounds from his already lanky frame, but for more than two months his blood glucose levels didn't budge. He began to fear that he would have to go on medication.

But finally, last November, Snyder's blood glucose levels returned to normal. According to his life insurance company, he's still diabetic — his condition possibly brought on as a result of his genetic predisposition and the added physiological stress of the viral infection that preceded his diagnosis. But so far he's managed to dodge the organ and tissue damage caused by excess blood sugar, and the side effects of medications. It's entirely possible that his iPOP, and its attendant 30 terabytes of data, saved his life. It also gave him new insight into how big data is likely to affect both doctors and patients.

"We've trained doctors to be so definitive in the way they treat patients," he says. "Every medical professional I encountered said there was no way I could have diabetes. But soon the volume of available data is going to overwhelm the ability of physicians to be gatekeepers of information. This will absolutely change how we do medicine." **SM**

Contact Krista Conger at kristac@stanford.edu

STATISTICALLY significant

BIOSTATISTICS IS BLOOMING

By Kristin Sainani

ILLUSTRATION BY JASON HOLLEY

When a baby girl, just days old, went into cardiac arrest in late January, her mother rushed her to the local fire station, where they defibrillated her and saved her life. She was sent to Oakland Children's Hospital and then to Lucile Packard Children's Hospital, where her heart stopped and had to be restarted multiple times. Her doctors put her on a heart-lung machine for a week, then implanted an internal defibrillator and removed nerve cells that make the heart jumpy. In all this turmoil, it's unlikely that anyone gave a thought to the role of biostatisticians in the girl's care. Yet, biostatistics played a part — helping to establish the effectiveness of the treatments that saved her daughter's life. And now biostatistics is playing a more overt role in her care. • Puzzled by the baby's condition, her doctors turned to Euan Ashley, MD, who is involved in several research projects to sequence the genomes of young patients with unexplained heart attacks. Sequencing a patient's genome now takes just three to four weeks — even faster than it takes to get the results of clinical tests for known genetic disorders, says Ashley, an assistant professor of cardiovascular medicine at Stanford. His team is combing through the 3 billion base pairs in the infant's genome, as well as those of her parents, to pinpoint what is likely a single genetic mutation responsible for her disease.

Sorting through the data isn't just a volume problem; rather, the data are inherently tricky. They contain tens of thousands of errors introduced by the sequencing technology. How to separate these decoys from the causative genetic change is an open challenge. "You're trying to sort out the needle from a stack of apparent needles," says Frederick Dewey, MD, a postdoctoral fellow in Ashley's lab.

This is just the kind of thorny data problem that is made to order for a biostatistician. Many types of expertise help crunch large data sets — computer science, mathematics and informatics, for example. But statistics brings something unique: Statisticians are not only trained in finding patterns in data, but also in separating real patterns from spurious ones. "Statisticians are very good at thinking about how bad their conclusions are," says Bradley Efron, PhD, professor of health research and policy and of statistics at Stanford and recipient of the 2005 National Medal of Science.

From personal genomes to gene expression arrays to electronic medical records, biomedicine is awash in tricky data. As a result, biostatisticians are increasingly in demand and in the limelight. For example, biostatistics was at the center of recent cover stories in both *The New York Times* and *The Wall Street Journal*. The phenomenon isn't unique to biostatistics; all of statistics is booming. As Google's chief economist Hal Varian, PhD, famously told a crowd at the Almaden Institute in 2008: "I've been telling people that the really sexy job in the 2010s is to be a statistician. Because they're the people who can make the data tell its story. And everybody has data."

Statisticians have their pick of jobs — Google, Facebook, pharmaceutical companies and tenure-track academic positions straight out of graduate school. "We're just not finding unemployed statisticians," says Ronald Wasserstein, PhD, executive director of the American Statistical Association.

"It's just a very exciting field," says Rob Tibshirani, PhD, professor of health research and policy and of statistics at Stanford and the second most cited mathematical scientist in history. "It's so much fun because people realize the need for statistics and they come with these very interesting problems."

The idea that statistics might be sexy, popular and fun may sound radical, but it's been gaining steam in biomedical circles for some time. Biostatistics has burst into attention recently, but only in the sense that a whale breaks the surface of the water, Wasserstein says. "The whale's been there all the time, and it's been having a huge impact."

Biostatisticians have changed the way doctors and biologists think, shaped the way they do research and built the tools for analyzing all types of data. Stanford statisticians have long been leaders in these endeavors; its statistics department is world-renowned and repeatedly ranks No. 1 on surveys of U.S. graduate schools.

Statisticians are also biomedicine's skeptics, Efron says. They scrutinize the evidence, and when it disagrees with conventional wisdom, they challenge the status quo.

BIOSTATISTICIANS MAY HAVE ENJOYED A LIFE of relative obscurity until recently, but their influence has rippled throughout medicine for nearly a century.

Statistical thinking revolutionized medicine by helping doctors focus on evidence rather than on intuition and feeling. "The general attitude that you ought to be quantitative and comparative in your thinking about medicine is a powerful idea that isn't natural to doctors. Or at least it wasn't from the Greeks until about 1930," Efron says.

Interpreting data is an art, because some of the associations that turn up are just flukes. Statisticians devised ways to separate these flukes from the truth. They also pioneered the concept of randomization — which makes it possible to compare two (or more) treatments fairly. "Before this, people blundered around for 2,000 years trying to decide whether A was better than B," Efron says.

In the 1970s, the randomized clinical trial became standard practice for evaluating therapies. Biostatisticians developed the methods for both designing and analyzing these studies. Efron was involved in several early trials at Stanford, including seminal studies by Saul Rosenberg, MD, and Henry Kaplan, MD, that established radiation treatment as a curative therapy for Hodgkin's lymphoma. "They changed it from an incurable disease to a curable disease," Efron says.

Biostatisticians continue to play a critical role in designing and analyzing clinical studies. "I keep telling people that researchers should not be analyzing their own data by themselves," says Helena Kraemer, PhD, emerita professor of psychiatry and

'STATISTICIANS are very good AT THINKING ABOUT HOW BAD THEIR CONCLUSIONS ARE.' THEY'RE TRAINED IN SEPARATING REAL PATTERNS FROM SPURIOUS ONES.

the go-to statistician in that department for over 50 years. "People who know what they want the data to say develop a certain functional blindness when it comes to what the data actually say." Statisticians can look at data honestly, since they don't have any stake in the outcome, except as patients, she says.

BIOSTATISTICS ENTERED A NEW PHASE in the 1990s and 2000s with the advent of high-throughput technologies — automated experiments that generate huge amounts of data, such as genome sequencing and microarrays. That's when bench scientists, who previously had little interest in statistics, began to recognize biostatistics as indispensable.

It wasn't immediately clear, even to statisticians, how to deal with so much data. For example, a microarray experiment might involve comparing the expression of 30,000 genes between cancer patients and controls. Traditional statistical tests allow one false positive to sneak in roughly every 20 comparisons; thus, when applied to microarray data, they generated a slew of false positives. Researchers excitedly proclaimed discoveries of gene signatures — patterns of gene activity predicting disease progression or treatment response — but the majority turned out to be nothing more than noise. Of hundreds of reported gene signatures, "almost none of them have panned out," Tibshirani says.

Statisticians soon became involved in debunking some of these baseless claims and in helping researchers sort the garbage from the real biological signal. Tibshirani developed some of the primary tools for controlling the false positive rate. His 2001 paper introducing SAM (significance analysis of microarrays) has been cited more than 7,000 times. This software program estimates the percentage of genes in a gene signature likely to be false positives and lets researchers cut the false positive rate by using more stringent criteria for gene selection.

He and other biostatisticians have also played important roles as "forensic statisticians." For example, in 2005, at the request of a colleague, Tibshirani scrutinized a high-profile *New England Journal of Medicine* article that claimed to have found a gene signature that predicts survival in follicular lymphoma. His conclusion after two weeks of work to reconstruct the analysis: The gene signature just didn't hold up. He wrote a letter to the editor laying out his criticisms and reanalysis, which the journal published. But the original paper continues to be cited — which illustrates a problem with the current publication process: Papers are often cited as fact even years after they've been discredited, Tibshirani says.

Recently, two biostatisticians — Keith Baggerly, PhD, and Kevin Coombes, PhD, of the MD Anderson Cancer Center in Texas — unearthed a scandal at Duke University that has led to the retraction of at least 10 papers in major medical journals. Their sleuthing revealed multiple bookkeeping errors and ultimately fraud in the data of the researcher at the center of the scandal. If it weren't for the persistence of Baggerly and Coombes, these problems, which were hidden within a complex data set, could easily have gone unnoticed — and clinical trial patients would still be receiving cancer drugs based on the fraudulent approach.

"As the data get more and more complex, it's easy to sort of massage it until you get a good answer," Tibshirani says. "As a result, biostatisticians are thinking about how to ensure that stuff that's published is more credible, trustworthy and reproducible."

As biology continues to evolve, so does biostatistics. Wing Wong, PhD, professor of health research and policy and of statistics at Stanford, is developing statistical tools for determining how genes work together in complex pathways in cells. "You cannot really understand the behavior of the cell by studying one gene at a time," he says. "There are some pretty deep statistical issues that we are struggling with."

CONTINUES ON PAGE 40

Need a biostatistician?

BIOSTATISTICIANS ARE IN SHORT SUPPLY. Fortunately, researchers at the Stanford School of Medicine can access one-on-one statistical consulting through Spectrum (the Stanford Center for Clinical and Translational Education and Research). Many universities provide similar services; and other researchers can access limited free statistical consulting on the web, such as at http://www.stat-help.com/.

"It used to be that people would come to statisticians at the end of the study, and we would do a post-mortem — tell them what went wrong," says Spectrum biostatistician Raymond Balise, PhD. But this has changed since the introduction of Spectrum's online study-design tool (http://spectrum.stanford.edu/studynavigator), which directs researchers to Balise and his colleagues early.

"There is a lot of basic advice that we can offer that gives studies more power, so researchers need fewer patients," Balise says. He also guides researchers in designing better questionnaires, planning for dropouts, measuring the right variables, and planning statistical analyses.

Spectrum's client load has nearly tripled since 2006, and the demand is "insatiable," Balise says. It's an incredibly rewarding job, he adds. "It's very hard not to fall in love with the research projects around here because they're all about saving lives and improving the quality of life," he says. "And biostatistics makes these projects possible."

KING OF THE MOUNTAIN

DIGGING DATA FOR A HEALTHIER WORLD

Medical researcher Atul Butte is at his computer collecting tissue samples, in this case for a study of leukemia. No need for sterile technique here — these days such tasks can be easily accomplished online. So he Googles. "Ah, here's a company in Huntsville, Alabama," he says. "OK, let's see: 'bladder cancer,' 'brain cancer,' 'breast cancer.' Ah, here's 'leukemia.' Let's click on that." • Butte, MD, PhD, an associate professor and chief of systems medicine in the Department of Pediatrics, has no lab in the orthodox sense. His discoveries, and there are plenty of them, pour out of a warren of cubicles housing computers and anywhere from 10 to 25 people. However untraditional, his lab is amazingly productive, averaging a new publication every two weeks — from new uses for old drugs to insights into the genetics of diabetes. • A few more clicks and he's ordered 15 serum samples from leukemia patients, and for comparison 15 serum samples from healthy people. "I get info on each patient's age, race, sex, alcohol and tobacco status, what medications they're on." And it's only $55 per sample. • Butte is visibly excited about this, as he is about many things. (He's a very happy man.)

By Bruce Goldman

PHOTOGRAPHY BY COLIN CLARK

AT LEFT: ATUL BUTTE AMONG THE DATA

His inflection rises. "They show up in 72 hours on dry ice. If we tried gathering them in my lab, it would take a year and they'd cost me about $1,000 apiece once I factor in all the labor that's going to be involved," says Butte.

Not that these samples are necessarily going to show up at Butte's doorstep. He'll probably outsource the analysis too — because it's faster, less expensive and in many cases is carried out with more expertise than he could muster even if he tried.

Traditional biological research has relied on painstaking work with experimental systems involving animal models: mice, rats, worms, frogs and flies, to name a few. But there's a new experimental system on campus. An explosion of biomedical data, particularly molecular data, is piling up exponentially in databases whose numbers are *themselves* increasing exponentially. The new experimental system is the universe of all these databases, many of which can be accessed and exploited via the Internet.

Butte is a walking window through which to watch the data revolution in bioscience unfold. A proven master at mining this medical data, he's now throwing his considerable energy into persuading other scientists to try his approach. It's the fastest, least costly, most effective path to improving people's health that he knows.

THE DIGITALIZATION OF BIOMEDICAL RESEARCH

Medical science is being swept up in a revolution, says Butte — a revolution as big as the microscope, or the breaking of the DNA code.

Although beginnings can be arbitrary, you could say this revolution — call it the digitalization of biomedical research — began with the Human Genome Project: a massive $3 billion effort, begun in 1990 and projected to take 15 years, to construct a linear catalog of all 3 billion chemical letters in a generic human genome. The Genome Project holds a hallowed place as one of the few large-scale government efforts that was completed early, in 2003. Only nine years later, we are fast approaching the era of the $1,000, 15-minute personal genome, a result of rapid technological improvements.

Butte thinks this means we now have to look at science in a new way. "In traditional biology research, people ask a key question, or run a trial. They make clinical and molecular measurements to address that question. They use some statistics or computation. Then they validate what they've found in another, more advanced trial. I would argue that three out of these four steps are now completely commoditized. We can outsource all that stuff and save a lot of money. But what you'll never outsource is asking good questions. As scientists, that's what we're really supposed to do best."

A lot of the answers to important medical questions are already here, Butte says, trapped inside a matrix of voluminous data gathering dust in myriad repositories — many of them accessible even to a teenager, and many of them free. The trick is to figure out what questions to ask to get the data to divulge their secrets.

"I don't think enough people study the measurements that have already been made," says Butte. "Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world. If I don't analyze those data and show others how to do it, too, I fear that no one will."

RIDING SHOTGUN ON AN AVALANCHE

Butte started writing code as a kid in New Jersey. He wrote it in longhand, inside notebooks. Tagging along with his parents on shopping trips, he would steal off to the computer-sales area and type his programs into the demo models. Eventually one of those shopping trips paid off big time. He got his very own computer at age 12 and never looked back. Butte picked Brown University for his undergraduate studies specifically because that school had an eight-year program allowing him to major in whatever he wanted (that was easy: computer science) then attend medical school there. On his own, the 43-year-old academic has spun off six companies during his career and is now spinning out his seventh. If you count the three more founded by his PhD students in the past year and not directly involving Butte, it's an even 10 so far. In 2011, he created the Department of Pediatrics' division of systems medicine, whose goal is to harvest the vast troves of biomedical data that researchers are pouring into public repositories. Thanks to the Internet, this data is now available to all comers (albeit with appropriate requisite safeguards).

The Human Genome Project was just a starting point. A genome, by itself, reveals only some of the mysteries of the organism. Biologists really want to know how many of which of our different proteins are at work in our cells, but proteins are tricky to keep tabs on. Because proteins vary radically in their structures and functions, quantifying them requires radically different biochemical procedures.

There's a proxy for that, though, stemming from the fact that genes carry the instructions for making proteins. In a living cell, the quantity of each type of protein being made can be estimated by measuring the quantities of a molecular intermediary, messenger RNA, that links individual genes to the proteins those genes specify just as a waiter's ticket links specific menu items to the plates of food the menu describes in words.

Different cell types in an organism "order" different amounts of the tens of thousands various human proteins, as does the same cell at various stages of development or de-

crepitude, or in disease versus health. So you can learn a lot about a cell's identity or condition by seeing what amounts of different proteins it's ordering up. This is known in the business as a gene-expression analysis.

Instruments pioneered by Stanford in the mid-1990s can simultaneously measure amounts of individual messenger RNA molecules (cells' order forms for specific proteins) corresponding to each of the roughly 23,000 genes in the cell's genome. These once-exotic devices, called microarray chips, are now a "throwaway commodity item," says Butte, retailing for as little as $200 apiece. The advent of these tools has triggered such a torrent of gene-expression experiments that "about a decade ago, the top peer-reviewed journals finally started saying, 'We won't publish your study unless you deposit its raw data in public repositories,'" he says. Government funding sources also now insist on such data dumps by grantees. The result: an explosion of data from studies employing genomics, gene-expression and other molecular methodologies. Thus, experimental data can be shared among not only the researcher and teammates but hundreds of other researchers they've never met or even heard of, and whose fresh approach may yield answers to questions that the guys who initially posted the data never thought to ask.

Health-related data can come from sources besides molecular measurements, including electronic health records, hospital orders, admissions and census data. Others are taking notice. Last year the McKinsey Global Institute noted in a report, *Big Data: The Next Frontier for Innovation, Competition and Productivity*, that analyzing the large data sets produced by the U.S. health-care system could create more than $300 billion in value a year.

GARBAGE IN, GARBAGE OUT?

Not everybody is in love with this new way of doing things. The sheer volume and novelty of data mining's output — a paper every 14 days! — is off-putting to more traditional bench scientists who earned their stripes mastering entire batteries of complex wet-lab techniques and overseeing or performing every step of the data-collection process themselves.

In April, at an off-campus retreat hosted by the Stanford Cancer Institute, Butte had an opportunity to sway doubters. He was accorded five minutes (the same as all the other speakers at the event) to talk about his methodology. There was little time for him to discuss his own studies in any detail, but most in the audience of about 100 were already at least broadly familiar with it. So, after a brief summary of all the major publicly accessible databases of possible interest to cancer researchers, he plunged into the question foremost on his mind. "OK, a show of hands," he said. "How many of you have actually taken advantage of these databases for your work?" About five hands went up.

Where's the data?

ATUL BUTTE MINES DATA STORED ON HARD DRIVES THROUGHOUT THE WORLD. HE SOCKS AWAY SOME OF HIS OWN LAB'S DATA ON THIS MACHINE IN ONE OF STANFORD'S DATA CENTERS.

"My second question is, why not? Your postdocs and grad students are perfectly capable of mining this data," Butte said. Up jumped the hand of distinguished cancer researcher Ronald Levy, MD, chief of the Division of Oncology. "Here is my answer to your question. Why should I trust data from experiments I haven't done, or overseen, myself?"

Butte's response: First, by definition, more than half of the data you pull from these databases is going to be quite recent. This follows directly from the fact that the contents of the biological databases are doubling every 15 months or so. So the data you download is going to have been derived, by and large, via up-to-the-minute techniques and instruments.

Second, Butte answered, "I can't vouch for the accuracy of the data in any given experiment pulled up from the database, any more than I can vouch for the accuracy of the

data from any specific researcher who's not a member of my own lab. But for the most part, the guys who are posting this data are the ones who are getting all the funding, so they tend to have histories of competence. And if their study was in a top-tiered government-sponsored database, they had [NCI director] Harold Varmus and [NIH director] Francis Collins looking over their shoulders. That's got to be worth something."

Did Butte's rebuttal change Levy's mind? "He made some good points, but I don't think it will convince many people to switch from doing their own experiments to relying on data from others who were doing *their* own experiments and asking their own specific question," says Levy.

And do these criticisms faze Butte? "Not a bit," he replies. "I probably have heard all the common criticisms. Successful scientists aren't necessarily going to change how they do science, but I do try to convince junior students and investigators that this data-driven approach is just as acceptable a way to build a career."

In any case, these database searches let you look at data aggregated from a huge number of independent experiments, not just from one lab. Just a week before the SCI retreat, Butte published a study in *Proceedings of the National Academy of Sciences* in which his group implicated a hitherto-unsuspected gene in type-2 diabetes. For the study, Butte checked out results from 130 independent experiments comparing gene activity levels in diabetic versus healthy tissue — in four tissues (fat, liver, muscle and insulin-producing pancreatic beta cells) and three species (mice, rats and humans). The same gene jumped over the moon in 78 out of the 130 experiments, a result whose chances of occurring randomly are less than one in 10 million-trillion.

Interestingly, the gene codes for a receptor found on the surfaces of macrophages, primitive immune cells that abound in porculent people's potbellies. An online search revealed that a famed experimental-animal facility, the Jackson Laboratory in Bar Harbor, Maine, had a strain of mice lacking the receptor. Butte ordered some of these mice along with their normal counterparts and had them sent out to an animal-testing consultant. Performing experiments designed by the Butte team, the consultant showed that while normal mice developed diabetes from a high-fat diet, otherwise identical mice lacking the receptor didn't. The team then tested a proto-drug blocking the receptor on the mice carrying the gene. That prevented these mice, too, from getting diabetes after waxing tubby on a high-fat diet. This proto-drug could turn out to be therapeutic for human patients, as well, and Stanford's Office of Technology Licensing is attempting to license the intellectual property related to the study.

It's discoveries like this that spur Butte to spread the gospel of data mining. In the past 12 months, he's given 28 invited talks in places as far-flung as London, Vienna and Seoul. It seems the word is getting out.

THE DATABASE MASH-UP

Butte says a 15-year-old can go to, say, the National Center for Biotechnology Information's GEO (Gene Expression Omnibus) database, type in "breast cancer," and get more than 31,000 experimental readouts — on more samples than any single breast-cancer researcher has ever put through a study — about as easily as chasing down a bunch of songs on iTunes.

JUST AS
ANY YOUTHFUL GEEK MIGHT CREATE
A MUSIC "MASH-UP," YOU CAN PAIR UP DATA SETS
AND SEE IF ANY INTERESTING CORRELATIONS JUMP OUT AT YOU.

Just as any youthful geek might create a music "mash-up" by syncing the vocal track of one song with the instrumental track of another, you can pair up data sets and see if any interesting correlations jump out at you. There are innumerable ways to match them up. You can cross-compare gene expression against blood chemistry (e.g., pollutant or vitamin levels), or census data against patient reports or patient care (what the diagnosis was, what was prescribed, what procedures were performed, what medications the patient bought and how consistently he or she took them, and patient outcomes), and tell the computer to let you know what correlates with what.

That's systems medicine, folks. And Butte *et al.*'s virtual-lab tests yield some very real-world results. For example, once you know which genes' activity is elevated or depressed

by a specific disease, and if in addition you know how those genes' activity is amped up or tamped down by various drugs, you can perform the molecular equivalent of a Match.com search, pairing drugs and disease indications according to the time-honored "opposites attract" principle.

In August 2011, Butte and his teammates published two papers in *Science Translational Medicine* describing how they showed just that. They used an algorithm designed in-house, which they made freely available to all researchers, hooking up drugs and diseases that had opposing effects on gene expression levels. The two papers detailed two separate such pairings: In one case, their algorithm predicted that a safe, old, off-patent ulcer drug, cimetidine, could be effective against lung adenocarcinoma, the most common form of lung cancer. In the other study, the digital Ouija board predicted that another safe, off-patent seizure drug, topiramate, could fight Crohn's disease, an inflammatory autoimmune condition affecting the intestinal tract.

But Butte and his associates didn't just stop at the formulation of a prediction. "At some point," he says, "you have to turn off the computer and actually try it." So they took the next step of testing their predictions in animal models. At first, they worked with other colleagues in the medical school who were equipped with animals and facilities. But to more rigorously verify their predictions, Butte wanted to avail himself of expertise not found on campus. For the topiramate study, he went online and found two companies that would conduct extensive trials in rats: induce a rat version of Crohn's disease, administer the drug, then analyze its effect by performing rat colonoscopies, captured on videotape.

That procedure, to say the least, requires technical sophistication. "No one at Stanford knew how to do rat colonoscopies," Butte says. Instead of picking just one company, Butte went with both, the better to demonstrate that the results, produced outside Stanford, were indeed reproducible. He clicked on "Add to Shopping Cart," provided his credit card information, and it was off to the rat races. The resulting video footage, as well as more routine histological analyses, showed that topiramate worked even better than steroids, which also have all sorts of potentially nasty side effects.

THERE'S GOLD IN THEM THAR HILLS

That study, which came out in August 2011, landed with a splash. "Essentially every major pharmaceutical company and biotech called within a month," says Butte. The academic community certainly took notice. "We're getting a couple of citations each week."

The federal government's research establishment noticed, too. "Historically, people have looked mainly at what a drug does, for good or ill, in a particular organ — the eye, the liver — and maybe not asked what this drug does in other organs. Nobody would have guessed topiramate was going to be useful in an inflammatory bowel disease," says Christopher Austin, MD, the director of the National Center for Advancing Translational Sciences' division of preclinical innovation at the National Institutes of Health. "Of course, it remains to be seen whether this is going work in humans. But a lot of the data Atul used to get his prediction was from humans. So these findings may turn out to be accurate."

Those who originally generated the data that Butte mined "had no idea what Atul was going to do with it all," Austin says. Butte's studies, he adds, show the importance of making the data public and "letting smart people like Atul, who had nothing to do with generating the data, follow up."

Long before the topiramate study was published, the molecular/malaise match-up algorithm was licensed by Stanford and became the core platform of a spin-off, NuMedii, cofounded in 2008 by Butte's graduate student Joel Dudley, PhD, and Butte's wife, Gini Deshpande, who has a PhD in molecular biology and biochemistry and whose background includes work for biotech companies and venture-capital firms.

The Mountain View, Calif., company "really took off last year," says Deshpande. "Atul and his group showed that old drugs can have surprising new uses. But we've been fielding calls from companies asking us if we can help them identify a *first* use for a drug they're developing." It's not always apparent which of perhaps four or five potential indications is the one in which a new drug is most likely to succeed, she says. And with the prospect of hundreds of millions of dollars going down the drain should a drug fail, the stakes in identifying the best drug/disease match are high.

Several other San Francisco Bay Area companies owe their existence to Butte's entrepreneurial instincts and data-mining discoveries. Among them is Carmenta Bioscience, a startup that blends database searches with observations of protein activities to address maternal and fetal health problems. Another, Personalis, aims to vastly improve the medical accuracy with which personal genomic test results are interpreted.

You might say Butte is fond of companies. He also likes company. In May 2012, Butte gave a talk to a group of Northern California science writers. What was billed as a 40-minute talk and Q&A became a tour de force that lasted well over an hour, at which point the restaurant started closing the place down. Forced out into the hallway, Butte cheerfully fielded waves of questions from a score of new fans, blissfully unaware that he'd left behind one of his props, a demonstration microarray chip, until someone noticed it on a table and brought it out to him. He shrugged, smiled, gave his thanks and said, "Did I mention this is a commodity item?" **SM**

Contact Bruce Goldman at goldmanb@stanford.edu

a singularity

AUTHOR VERNOR VINGE TALKS SCI-FI HEALTH CARE WITH
SCHOOL OF MEDICINE CHIEF COMMUNICATIONS OFFICER PAUL COSTELLO

I missed the *Star Trek* era. *2001: A Space Odyssey* didn't really appeal to me either. And Robert Heinlein novels? Sorry. So I approached an interview with celebrated science fiction writer Vernor Vinge with a bit of trepidation. Would we have a common language, or would I end up lost in space?

VINGE, AN EMERITUS PROFESSOR OF MATHEMATICS AT SAN DIEGO STATE UNIVERSITY, is considered one of the greatest science fiction writers alive today. A five-time winner of the Hugo Award, science fiction's most prestigious honor, his stories explore themes including deep space, the future, and the singularity, a term he famously coined for the future emergence of a greater-than-human intelligence brought about by the advance of technology.

But why interview Vinge for a medical magazine? Well his track record for prognostication is pretty good. In *Rainbows End*, his 2006 novel, he paints a world of augmented reality where computer and wireless technology converge with people's eyewear so life becomes an interaction with data that never ends. Sounds a bit far-fetched but it's not: Google's Project Glasses hopes to launch just such a product in the near term. So we thought we'd take a leap and have Vinge consider health care, medicine and big data — all worlds that he's touched on in his work. Will he be prescient here too?

PAUL COSTELLO: In your novel *Rainbows End*, you create a world where vast stores of data are accessible at all times — literally in your face, seen through contact lenses. If this is a realistic picture of our future, what do you see as the consequences?

VERNOR VINGE: What we have is data glut. What we really want is the ability to manipulate the information and to reach conclusions from it. I think we are at the point where that is slipping beyond unaided humans' abilities. So the real thing to be looking for is processing schemes. One way is automatic processing: for instance, the sort of analysis that we saw with the IBM Watson on *Jeopardy*. Putting that in service to humankind in fields that are suffering from data glut at least gives people who are in charge the ability to keep some sort of track of what is going on.

The other great thing that we have going for us is that we have billions of very intelligent people out there in the world. With the networking that we have now, we're beginning to see that those large populations, coordinating amongst themselves, are an intellectual resource that trumps all institutional intellectual resources and has a real possibility, if it's supported by the proper automation, of creating solutions to problems, including the problem of the data glut.

COSTELLO: Your Internet presence is pretty minimal. No Facebook presence to speak of, no Twitter. You seem to value your privacy. But you see social networks as the future?

VINGE: Yes, I value my privacy. I think that's a matter of personality more than anything else. I think that we may be seeing something

sensation

here that is close to being something new under the sun, or at least, it's a quantitative change in sociality that is so large it actually could count as a qualitative change in human nature.

There was a big step in that direction with the rise of humanity. The evidence is that some of our biggest jumps occurred when you had stable societies where you could have apprentices. And so you could get a critical mass of innovation. I think we are going through another critical mass of innovation now with things that are sort of flippantly called "crowdsourcing."

Now, at the same time, there may be people who do not do very well with that and do not participate very well in that. I think many of those are at a substantial disadvantage, and I don't discount myself in that regard.

COSTELLO: With all this information at the tip of our fingers, how does this change health care? In this world, would we even *need* physicians?!

VINGE: I think that we will need them. I think one thing that we'll see is the ability to get your medical information, what is going on inside your body, in real time on a second-by-second basis. Talk about data glut! That would mean that a lot of acute causes of death that exist right now would be relatively easy to head off. There would be doctors and other medical professionals in charge of tracking these sorts of things and managing interventions that might occur well ahead of the sort of catastrophic failures that now come as such a big surprise to people, say when they have a stroke or a heart attack.

COSTELLO: Medical technology gave the hero of *Rainbows End* a new lease on life, making him young again. Is that the pot of gold at the end of the rainbow? Young again?

VINGE: Ah! Actually, I think one scenario to really think about is the one where we don't exactly get immortality, but what we do get is resilience — that the resilience of youth can exist at any age. So you still have various diseases, but the resilience and the grow-back possibilities remain high indefinitely.

And I think there's some possibility of that version of "prolongevity" happening, and if it did, all sorts of very, very interesting and surprising things would happen. It would probably have a social effect as great as any political revolution. Because right now, the old people have most of the money, but they don't have any get up and go. So if your average 80-year-old had the get up and go and mental alertness of the average 50-year-old nowadays, they would kick ass. You could very quickly get into a situation where society would essentially turn upside down, and if you were under 50, you would probably be peeved, although there would be the prospect that you would, eventually, do better.

I think that, in time, being able to have the advice of people who were 200 or 300 years old and have seen a lot would have a very good effect on the overall stability and survivability of the human race.

COSTELLO: What medical advance are you most looking forward to?

VINGE: Well, obviously that resilience of youth would probably be at the top of the list, but the advance that's doable possibly in the very near future would be memory improvement, enhancement drugs. Having a sharp memory would benefit an enormous number of people, and, very selfishly, me included.

This interview was condensed and edited by Rosanne Spector.

RICCARDO VECCHIO

WEB EXTRA HEAR THE INTERVIEW AT HTTP://STAN.MD/NgzPkV

on the records

TAPPING INTO STANFORD'S MOTHER LODE OF CLINICAL INFORMATION

By Erin Digitale

ILLUSTRATION BY SHANNON MAY

On a spring evening in 2000, Henry Lowe, MD, had just finished dinner at a downtown Palo Alto restaurant with a group of Stanford biomedical informatics students. A master at assembling large, complex databases from multiple sources of patient data, Lowe was in town to interview for a position on the faculty. As they strolled under the plane trees along University Avenue, one student took him by surprise, asking: "Are you sure you want to come to Stanford?" "Yes, I think I do," Lowe remembers saying. • "Well, it's nearly impossible to get access to any clinical data for research," the student said. "When our researchers want clinical data, they have to go to Kaiser or the VA." • Telling the story now, Lowe pauses, choosing his words.

"I came to Stanford with the goal of changing that," he says.

It wouldn't be easy. "Back then, the idea that I was going to go to the hospitals and say, 'Give us all your clinical data,' seemed absurd," Lowe says. "Hospitals are, for all the right reasons, very sensitive about protecting that data."

Not only did Stanford's two hospitals — Lucile Packard Children's Hospital and Stanford Hospital & Clinics — lack a tradition of using their patient records for research, but the records were housed in several unrelated databases. Organizing the information would require a major investment of time and money, and new technical and legal structures to accommodate re-use

of the data. Yet other institutions — including the National Institutes of Health and a few academic medical centers — were establishing such systems. Despite the obstacles, Lowe felt a Stanford clinical data warehouse was essential. "We have an amazing opportunity to use our patient-care-delivery system as a sort of natural experiment," he says.

"If you look at the FDA statistics, the time required to get an idea to bedside is years," says Elaine Ayres, deputy chief of the Laboratory for Informatics Development at the NIH Clinical Center, which has a clinical research data repository of its own. "These systems have the ability to shorten that cycle." The rich lode of clinical data allows researchers to tackle questions that are logistically and ethically impossible to address with clinical trials, and enables much larger studies than traditional clinical trials. As demands for evidence-based medicine increase, the data provide a tremendous evidence base to tap.

Today Stanford is a national leader in the use of clinical databases for research. Lowe, senior associate dean for information resources and technology at the School of Medicine, spearheaded the development of a world-class clinical data warehouse at Stanford — the Stanford Translational Research Integrated Database Environment, known as STRIDE.

The fast, intuitive searches the database makes possible are engendering a new kind of scientific creativity at Stanford, as researchers preview the contours of Stanford's population data — more than a million patient records and growing — to help brainstorm hypotheses and develop studies. "People look at this and they salivate over it," says Lowe, who also directs the Stanford Center for Clinical Informatics. Scientists from as far away as Australia, New Zealand, China, Korea, Norway, Finland, Great Britain and Ireland have come to learn how STRIDE was built. He's happy to share Stanford's know-how, though he's the first to admit that even after a decade, in many ways the project is still a work in progress.

When Lowe arrived at Stanford in 2001, the School of Medicine had an ambitious new dean, Philip Pizzo, MD, who wanted to promote translation of basic discoveries into clinical care. Pizzo was fresh from a stint at NIH, which led the world in tapping electronic medical records for research. His enthusiastic advocacy for a clinical data warehouse for research was integral to the success of Lowe's effort.

Pizzo helped convince leaders of the two Stanford-affiliated hospitals that a research database based on patient records would give the entire medical center a powerful advantage. In addition to the ability to advance medical science, the hospitals would gain an excellent platform for assessing patient safety and quality of care. Pizzo also committed to financing STRIDE's development and operation, at an eventual cost of several million dollars.

Todd Ferris, MD, now director of informatics services at the Center for Clinical Informatics, soon joined Lowe's team to help determine how to protect the privacy of patients in an ever-growing pool of live clinical data while providing scientists access for research. The team spent a year and a half crafting an agreement with the hospitals, then subjected their plan to an external legal review.

With administrative, financial and legal backing in place, Lowe and his colleagues tackled logistics. In 2003, after a fruitless hunt for a database that could integrate the many types of data in clinical records — such as patients' diagnoses, demographic information, prescriptions, lab values, radiology

images and pathology reports — they decided to design and build a database instead.

"Slogging through the data was painful and time-consuming but it also engendered a huge amount of knowledge for the team," Ferris says. "If we had bought a database and pointed a pipe at it, we might not have developed the same level of understanding."

"We were acquiring data from dozens and dozens of clinical systems," Lowe adds. "Then we had the challenge of unifying all of it, figuring out which patient each piece belonged to, and assembling it in the database."

The sheer volume of data was part of the problem. STRIDE now contains 1.7 million patient records. "We are a small medical system in comparison to some others," Lowe says. "But that's still a very large data set."

The team developed a wish list of functions for their database, many of which are now incorporated into the slick user interface that Stanford researchers use to access STRIDE, which went live in 2005.

"Because this database is comprehensive, centralized and continuously adding new information, it provides a unique and visionary resource to our community," Pizzo says. "STRIDE anticipated the incredible changes now occurring as medicine becomes increasingly quantitative."

After almost a decade of development, STRIDE is ready for the big time. It's on par with the best similar databases in the country at institutions such as the Mayo Clinic, Harvard/Partners Healthcare and Vanderbilt, says the NIH's Ayres. A few large medical systems, such as Kaiser Permanente and the Veterans Affairs Administration, also have regional databases, but these systems so far lack the comprehensive capabilities of those at large academic medical centers. When Lowe dreams really big, he thinks about how to link STRIDE data to medical records from other institutions across the country. For starters, his team is working on meshing STRIDE with databases from other health-care systems in Northern California. It's daunting because the total data set is likely "1,000 times bigger and more diverse" than STRIDE, he says.

Today, when a researcher wants to determine if STRIDE contains the right data for a particular research project, the first stop is the system's Cohort Discovery Tool, built to quickly query medical records without revealing the identity of any individual patient. "Let's look for a group of patients that could be the basis for a study," says Lowe, as he logs in to the online STRIDE interface. "How about young heart attack sufferers?" He drags "Diagnosis" from a menu on the left side of the Cohort Discovery Tool's screen to the work space in the center, typing "myocardial infarction" in the search box that pops up.

Behind the scenes, the system matches his query to ICD-9 diagnosis codes, the standardized descriptors used in medical records. Hitting the "Go" button, he sees that about 9,625 Stanford patients have had an acute myocardial infarction.

"Let's say we're interested in clinical events from just the last five years," Lowe says. He chooses "event date is after" and enters "01/01/2007." The system spits out a new number: 3,350 patients. Narrowing again, he specifies "age at event is equal to or less than 55 years." About 685 patients meet the narrowed criteria. In less than two minutes, he has assessed the electronic medical record data in a way that would be almost impossible by hand. Now he can further refine his search — by specifying patients taking particular medications, for instance, or those who have a certain type of lab result on file. As he searches, the system automatically displays bar graphs of the cohort's demographics: gender, current age, race and most recent address.

For scientists building a study, the next step is a data review. Following a formal review by the school's privacy officer, researchers may obtain permission to review data from all patients in the cohort to see whether their records will really answer the questions the researcher wants to ask. "There's an awful lot of noise in clinical data; you can't assume the answer from the Cohort Discovery Tool is correct," Lowe says.

To conduct a study using STRIDE data, researchers need to get approval from the Institutional Review Board — the organization that gives the thumbs up or down to any research on campus involving human subjects. As required by federal privacy statutes, the IRB assesses the minimum amount of information necessary for scientists to answer their research question. Many studies can be conducted with de-identified data — stripped by STRIDE of the patient's name and anything else that could be traced back to the individual. Other studies require identified data. In some such cases, patients are contacted individually to request their permission to include them in the study. However, as with any proposed study of human subjects, if the overall risk to patients is judged to be negligible, the IRB sometimes grants a waiver that permits scientists to proceed without contacting individual patients first.

The IRB is the key patient-privacy gatekeeper, Lowe says, emphasizing that the IRB has the ultimate control in ensuring that researchers get only the data they need and no more. STRIDE also has several methods of ensuring that its interfaces cannot be used to underhandedly triangulate back to a specific patient. Another privacy safeguard is software that enables researchers to analyze patient data without downloading it from STRIDE's secure servers.

The Notice of Privacy Practices that Stanford and Packard Children's patients receive informs them about potential re-use of clinical data for research, and patients have the op-

tion to request in writing that their data never be used. Yet Lowe hopes patients will see the big-picture value of this type of research, adding "Here's an opportunity for your data to be a small part of a much bigger data set that could lead to dramatic discoveries and breakthroughs in understanding."

To help scientists build good studies, in 2005 Lowe's team inaugurated the Stanford Center for Clinical Informatics. Demand for its services is steadily increasing: They received 318 requests for "substantial informatics consultations" to help scientists use STRIDE in 2011, more than double the requests per year in 2008 and 2009, and up from 198 in 2010. Clinical informatics research is catching on at Stanford, says Lowe, though he still thinks the gold mine of information in STRIDE's data has barely been touched. So far, only a handful of studies have been published using STRIDE data.

LOWE AND HIS COLLEAGUES ARE NOW TACKLING WHAT HE CALLS 'THE HARD PROBLEM IN MEDICAL INFORMATICS': EXTRACTING USEFUL INFORMATION FROM TYPED OR DICTATED TEXT WITHIN MEDICAL RECORDS.

Among STRIDE's converts is Russ Altman, MD, PhD, whose team used the database to explore adverse drug interactions, turning up several. • "Initially, many of us were very nervous that medical record data would not be good for research because it would be so biased by the purposes for which it was designed," says Altman, a professor of bioengineering, of genetics and of medicine. Although electronic medical records are indeed biased, for instance by clinicians' tendency to favor diagnosis codes with higher insurance reimbursement rates, Altman says that with careful planning to deal with such bias, "I believe they're extremely valuable."

One of Altman's studies, published this year in *Science Translational Medicine*, revealed potentially dangerous interactions between widely used drugs by tapping both STRIDE and the Food and Drug Administration's database of more than 4 million adverse-event reports collected across the country.

"It's only with the emergence of very large public databases that you can begin to ask questions about drug-drug interactions," he says.

The scientists designed an algorithm that trolled through the FDA's adverse-event data seeking pairs of patients who are extremely similar except for one drug prescription. Comparing these matched patients made it easy to spot side effects produced by drug combinations. The research found several dozen interactions; the biggest discovery was that patients taking both SSRI antidepressants and the thiazide type of blood pressure medication are at increased risk for a potentially deadly form of heart arrhythmia, long QT syndrome. The team used STRIDE data to check their findings. Sure enough, patients in STRIDE who were taking SSRIs and thiazides had increased risk of the arrhythmia.

One advantage to big data: "We're slowly learning that you can throw away data that's not perfectly useful," Altman says. For instance, the adverse-event study used only data from well-matched pairs of patients, ignoring data from patients who could not be matched according to the algorithm his team designed. "Traditionally, statisticians weren't able to insist on that because they didn't have big enough data sets."

But doesn't throwing out data open the possibility of bias in what is kept? Not really, says Altman — because of the sheer size of the data sets. Unlike the findings from a traditional study, these numbers won't fit into a spreadsheet. There is no way for scientists to see the numbers before they decide how to analyze them. As Altman puts it, "It would be very hard to cherry-pick the data even if we wanted to."

Lowe and his colleagues are now tackling what he calls "the hard problem in medical informatics": extracting useful information from typed or dictated text within medical records. This data trove remains difficult to tap because the computational challenge is far more complex than processing the structured portion of the clinical record.

There's an irony in the situation: A big motivation for creating electronic medical records was to compute their data, but so far that's extremely difficult for the richest, unstructured, portion of the information.

"So the quest continues," says Lowe. "We have the electronic health record, but we still can't compute much of the really important data it contains."

But they're making progress. The text-mining tool they've begun developing can already parse pathology reports to find patients with a specific cancer diagnosis. And it can pull out descriptions of cancer types and biopsy sites. It even has rudimentary "negation detection," the ability to understand a notation that the patient does *not* have the diagnosis of interest.

Still, the tool is much less powerful than a human reader. But just give it a little time, says Lowe. **SM**

Contact Erin Digitale at digitale@stanford.edu

GAME ON

STANFORD DEVELOPS A NEW TOOL FOR TEACHING DOCTORS TO TREAT SEPSIS

By Sara Wykes

PHOTOGRAPHY BY COLIN CLARK

Jack was sinking fast,
his vital signs registering alarming numbers.
With every passing second,
his doctor,
Charles Prober,
could see his patient being overwhelmed by sepsis,
a deadly complication of infection that plagues hospitals worldwide.

"Jack is the hardest patient,"
counsels Prober's colleague, Lisa Shieh, MD, PhD,
the medical director of quality in the Department of Medicine at Stanford Hospital & Clinics.
"Give him some fluids."

Prober, MD, the senior associate dean for education and a professor of pediatrics and of microbiology and immunology at the School of Medicine, clicks in the order.
A small group of watching physicians clapped in appreciation as Jack's health almost immediately improved. • "Has he had his blood cultured yet?" asks Shieh. Prober takes the cue. Then he turns to another patient ailing from sepsis. Just as he finishes ordering fluids for that one, his colleagues shout, in alarmed voices, "Jack, Jack!" — the first patient's previous gains were rapidly evaporating. "Ah, that's one thing we want to teach," Shieh says. "You can't just give fluids and walk away." • Prober responds quickly, transferring Jack to the intensive care unit and setting up a surgical procedure to remove infected tissue in Jack's leg, among other steps. Jack's status zooms to complete health: Prober is awarded 500 points. • Prober has scored well on a test run of a first-of-its-kind, medical computer game called Septris (named after the popular game Tetris). The idea is to tap into the power of games — an increasingly popular technique, called gamification — to improve clinicians' skills at recognizing and treating sepsis. Created by Shieh and a team of Stanford physicians, researchers and education technology experts, the game can be played on a mobile phone, a tablet such as an iPad, a laptop or desktop computer • In real life, sepsis begins as a bacterial infection at a single source, which, if uncontrolled, spreads to

LISA SHIEH (LEFT) AND EILEEN PUMMER LED A GROUP THAT CREATED SEPTRIS, A MEDICAL COMPUTER GAME THAT'S IN PLAY ON THE SCREEN BEHIND THEM.


Libby
Back
Next

become a systemic attack on the body's kidneys, liver, lungs and central nervous system. It presents as simple sepsis, then moves to severe sepsis and, finally, to septic shock. It can run its entire course within hours. Unless it's stopped at its earliest stage, sepsis can claim one life in every two it invades. More than 200,000 Americans died last year of sepsis. In the last year, the cost of care for the disease amounted to $2 billion in the United States.

"Sepsis is one of those conditions you hear about in med school, but you need to see more of it," says Shieh, a clinical associate professor of medicine and director of the group that developed Septris — the Stanford Hospital sepsis performance improvement team. "In some cases, it's straightforward, and in some cases, it's not. It takes a lot of clinical sense."

'SEPSIS IS ONE OF THOSE CONDITIONS YOU HEAR ABOUT IN MED SCHOOL, BUT YOU NEED TO SEE MORE OF IT.'

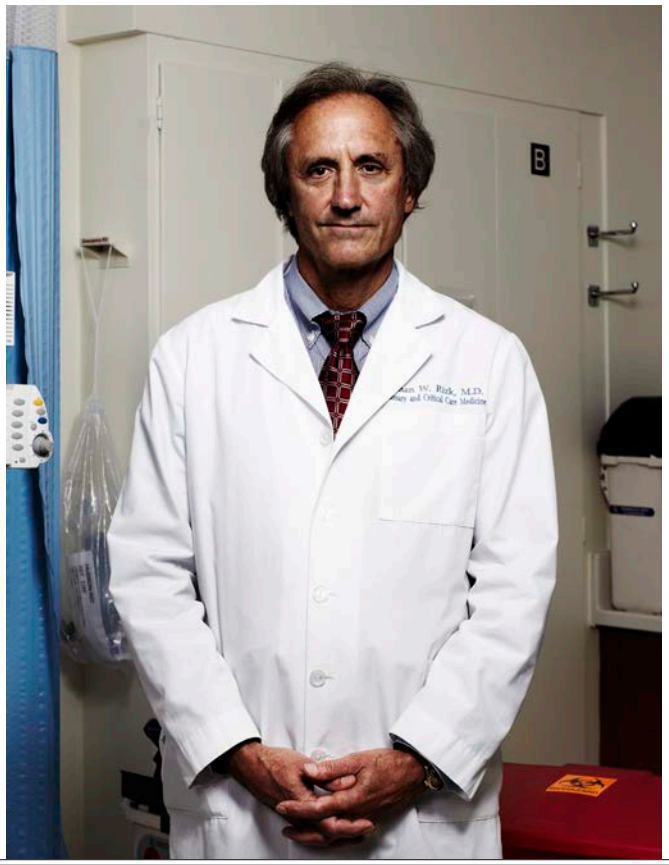
The game begins with the cartoon image of two patients on the left side of the screen. On the right side are their vital signs. Along the bottom of the screen are diagnostic tests and treatment options. As every second passes, the patients' images sink down the screen, their vitals deteriorating. It takes less than two minutes for a Septris patient to die, which means decisions must be made quickly. The game's objective is not just to keep the patients alive, but to cure them.

COEXISTING CHRONIC CONDITIONS COMPLICATE DIAGNOSIS OF SEPSIS: They can make a patient more vulnerable to sepsis, but also distract a doctor from identifying it. "Everybody needs to have at least the Septris level of sepsis knowledge," says Norman Rizk, MD, medical director of the hospital's intensive care units and professor of pulmonary and critical care medicine. "This simple training tool begins to establish essential knowledge."

Septris is the result of a $30,000 grant from the medical school's continuing medical education program to develop a more effective way to teach how to treat sepsis. The team first considered a lecture or workshop format, the typical formats for CME courses. But by the team's second meeting, people were looking at their phones and losing interest, says Eileen Pummer, RN, a quality manager at Stanford Hospital and the team's co-director. "I thought, 'Oh, no, this is falling apart already.' Then Matt Strehlow said something to the effect of, 'How about a mobile app? People are always on their phones.' The energy just turned completely around, and we started brainstorming from there."

Strehlow, MD, is a clinical assistant professor of surgery and an emergency medicine physician. He is also the assistant medical director of the hospital's emergency medicine department, with a special interest in educational technology and in sepsis. Strehlow is so addicted to computer games that he doesn't have any on his phone or at home to avoid the disruption it would cause to his life. He's very aware of the popularity of computer games and of how many medical students and young physicians use their iPhones and iPads as knowledge support tools. During a postdoctoral study in India, Strehlow tested the teaching power of gaming against more traditional simulation methods and reading. In initial comparisons, the simulator-learners performed better than did the gamers. Three months later, however, he retested the group and found the gamers ahead on skills. "They'd been going back into the lab to play the game," he says.

The clinicians brought the idea to educational technology manager Brian Tobin and instructional technologist Jamie Tsui in the medical school's Office of Information Resources & Technology. While the team had solid content — best practices and guidelines based on medical literature — there were different questions to be answered for a game format. One of the biggest, Tsui says, was whether to allow the game patients to die. "Players had to be allowed to fail, but also to have a chance to fix their mistakes," he says. So, patients die once and then reappear with the same symptoms, thus providing the opportunity for knowledge acquired by failure to be applied with success the second time through. The de-



STANFORD HOSPITAL ICU CHIEF NORMAN RIZK IS A STRONG BELIEVER IN PLAYING SEPTRIS TO GET BETTER AT SAVING PATIENTS FROM SEPSIS.

signers also limited the number of patients in play at a time to two, though Tsui noted that more could be added in the future if users are "up for that higher level of play."

The goal was to keep fun in the experience, despite the gravity of the topic. "You want to let the learner create and play, and you want to offer them choices," says Tobin, who's now acting director of educational technology. "It also has to be suitably hard enough so that not anybody can get right through it."

"At first, we made it way too hard and patients were dying too fast," Strehlow says. That first version had eight patients on screen at the same time. The group tested the game on several groups of physicians to work out the combination of symptoms and timing that would be challenging without being impossible to beat. There's also a classic trick: At least one of the patients doesn't have sepsis, Strehlow says.

MEET DR. SEPSIS

Making the right choices when you're playing Septris leads not only to more points but to a visit from Dr. Sepsis, a character who pops up on screen offering praise and pointers. That he bears a striking resemblance to ICU chief Norman Rizk, MD, is no coincidence.

They adjusted the speed of the game to accommodate clinicians who "like to take their time reading test results versus those who work very quickly," Tsui says. They did not include images of CT scans or X-rays, however. "That added a layer of complexity to a game we wanted to keep as simple as possible," he says.

"We wanted to build something that would work across all platforms," says Tobin, "whether someone is using a handheld with a touchscreen, or a computer, where you can just click a link and the game displays right then and there."

The Septris team had to make medical decisions, too. In real life, some antibiotics work better in combination with others; the game awards points if a player understands the possible positive or negative effects of those combinations. They also wanted to reward players with more than points for making good choices. When a good choice is made, a pop-up appears with words of praise and wisdom from "Dr. Sepsis," whose knowledge-reinforcing tips are meant to be like those from an attending physician to a less experienced physician. The game's tips section has links to medical journal articles that back up the practice guidelines Septris teaches.

The game is not without a bit of subterranean Stanford medical community humor. The cartoon figure of Dr. Sepsis looks a lot like Rizk, the ICU chief and senior associate dean for clinical affairs. And the pretend patients are quite reminiscent of some other Stanford physicians.

TRY IT

Playing Septris is free. It runs best on iPad/iPhone or Android. On a desktop computer, it requires a Firefox, Google Chrome or Apple Safari browser. To start the game and to learn about CME credits that are available upon completion, go to http://cme.stanford.edu/septris. The $20 fee for the test for the CME certificate is waived for the first 100 learners.

The Septris team knows that some people are uncomfortable with using a game to teach such a serious subject. Even some computer gaming fans were a bit troubled by it. "The comment was that there's something about the word 'game' that doesn't feel like the right fit when you're thinking about treating patients," says Clarence Braddock, MD, associate dean for undergraduate and graduate medical education and a professor of medicine.

But Septris is not conceptually different from the modes of simulation now being used at Stanford and other medical schools to train physicians. "What you're talking about are ways to activate a learner's mind to engage and connect," Braddock says. "You're trying to mimic the cognitive pressures and drilling around the application of concepts to clinical problems."

In fact, the team members think the approach could teach many medical topics: This summer they'll start working with Stanford surgeons to use the same platform to teach about surgery cases.

Braddock and other medical school educators plan to study the game's effectivenesss. The first large-scale group of subjects will be this summer's incoming interns.

Already, though, the game has one benefit: People like playing it. Since its release this winter, it's been played 10,000 times and had 14,000 unique visitors from 54 different countries, more than 1,000 each from Brazil and Australia. Charlie, the patient who's easiest to help, has been saved 2,300 times; Jack, who's the toughest, has died 1,400 times.

Anyone can play Septris: It's available at http://cme.stanford.edu/septris/, and it's free. Physicians can earn continuing medical education credits by taking a post-game test.

Interest in the game is still growing. Shieh has been invited to speak about it in July at a world conference on using mobile devices in medicine. Officials at the Wake Forest School of Medicine have also asked Stanford's School of Medicine to set up a one-week competition between the two. "They said they found the game really intriguing," Shieh says.

"Septris puts people through an increasingly more realistic and challenging simulation and generates adrenaline by having fun and engaging in it," Braddock adds. "You're not thinking of it as learning, but as play, which, from a neurochemical standpoint, is a win-win." **SM**

Contact Sara Wykes at swykes@stanfordmed.org

cancer

ROUNDHOUSE

By Christopher Vaughan

ILLUSTRATION BY JEFFREY FISHER

"It's no fun to manage a patient's death," says Stanford physician and researcher Ravi Majeti, MD, PhD, with grim and deliberate understatement. As a physician specializing in treating leukemia and lymphoma, Majeti is often required to perform that task with as much compassion and empathy as possible. In the case of acute blood cancers, even in young patients who undergo aggressive treatment, the odds of surviving more than five years is less than 50-50. For those over age 65, Majeti says, five-year survival rates are about 10 percent.

What's particularly frustrating is that these atrocious odds haven't changed in decades, says Majeti, an assistant professor of hematology and a member of the Stanford Institute for Stem Cell Biology and Regenerative Medicine and the Stanford Cancer Institute. "The primary drugs and the combinations we use have not changed in 30 years. They *have not*," he says. If a patient fails a first round of chemotherapy, it is difficult to halt the malignant blood stem cells from multiplying uncontrollably, crowding out nearly all the normal blood cells that protect us from infection, carry oxygen and clot to keep us from hemorrhaging. Lung infections are usually the ultimate cause of death, he says.

In large part because of his frustration with these terrible and unchanging statistics, Majeti spends much of his time in the laboratory, looking for new ways to shift the odds in the patients' favor. Some of his work involves taking samples of leukemia cells from patients in the hospital and putting them into mice to test new therapies.

So it was with some small hope, on a day in November 2007, that Majeti tested a new antibody discovered in the laboratory of institute director Irving Weissman, MD. Majeti and MD/PhD student Mark Chao first injected a mouse with aggressive human acute myelogenous leukemia cells. This particular leukemia had, in fact, eventually killed the patient who donated a blood sample for research. Injected into a mouse, leukemic blood cells normally do the same thing they do in humans — multiply out of control until they kill the host. What happened next was one of the most astonishing moments in Majeti's scientific career.

"For the first test, we were just guessing a dosage and hoping we could observe some small effect," Majeti says. But there was no small effect — it was huge. A day after injecting the antibody, Majeti and Chao couldn't observe any cancer cells in the mouse at all. "One dose, one day, and the cancer was gone," Majeti says. At first Chao thought he had done something wrong, but there was no mistake. The same lethal cancer cells that could not be stopped in the human patient, the same cancer that was so irritatingly resistant to everything the doctors could throw at it in the clinic, had simply disappeared in the mouse after being exposed to a single dose of experimental antibody.

As dramatic as that experiment was, further research kept producing new amazements, suggesting that the applications of this antibody to cancer therapy are far broader and more powerful than anyone dared hope. The experimental antibody that Weissman and his collaborators discovered blocks a cell protein called CD47 — a cellular cloaking device that offers cancer safe passage from immune cells that eat damaged cells or foreign matter.

Investigation into the role of CD47 began slowly 14 years ago in Weissman's laboratory, but like a snowball kicked off a hilltop, it has picked up speed and mass as it has rolled along. It now seems on a course to blast through the traditional cancer treatment community. As the researchers prepare to test the treatment in humans, they have dared to hope that they're on the trail of something many have dreamed about but most had begun to think impossible: a single therapy that uses our own immune system to effectively attack all cancers with almost no side effects.

THE "DON'T EAT ME" SIGNAL

It is the nature of life that things will go wrong eventually. Our cellular software, our DNA, can get damaged in many ways. Eventually "bugs" in that software accumulate, and cells stop following instructions written and revised over billions of years to make sure they do their proper jobs. One result of this process is cancer — cells that are supposed to behave within the rules of the body's decorum begin breaking those rules and multiplying out of control.

In 1998, Weissman and his postdoc David Traver, PhD, were crossbreeding mice with various genes that block programmed suicide in cells that have been damaged, genes that are known to be associated with cancer. They created a breed that was particularly prone to developing leukemia, then analyzed all the genes being manufactured (or "expressed") by the blood-forming cells in these mice. "The first gene that we saw that was overexpressed in the mice that got leukemia was CD47," Weissman says. In fact, it turned out that high levels of CD47 were common in every kind of leukemia, in mice and humans both.

But no one knew what CD47 did. Then, two years later, a group in Sweden discovered that one role of the CD47 protein was to act as an age marker on red blood cells. They discovered that red blood cells start out with a lot of CD47 on their cell surface and slowly lose the protein as they age. At a certain level, the dearth of CD47 allows macrophages to eat the aging red blood cells, thus making way for younger red blood cells and a refreshed blood supply. CD47 thereafter became known as a "don't eat me" signal to the macrophages.

For Weissman, the Swedish work and the work in his lab suggested an explanation for leukemia cells' invincibility. Leukemia, which is a disease of excess production of specific

sorts of blood cells, always features high levels of CD47. And CD47 naturally protects red blood cells from being cleared away by the immune system. Could blood cancer cells be boosting levels of CD47 to protect themselves from being consumed by macrophages?

Oddly enough, no one in Weissman's lab was interested in looking for an answer to that question. Principal investigators with large labs generally don't conduct experiments themselves, relying instead on an army of postdocs and students to do research under their direction. Weissman, unlike many principle investigators, doesn't order people in his lab to carry out specific research, preferring instead to let them pick projects that interest them. "I've found over my career that the people I had to direct more closely never developed as scientists," Weissman says. "Whereas those people who take on problems and figure out how to approach them go on to become accomplished scientists."

So for three years, Weissman says, he found himself saying, "Come on you guys, how could it be more apparent? Every mouse leukemia has CD47, and CD47 is a 'don't eat me' signal to macrophages. It has got to be important."

Finally, in 2003, an MD/PhD student named Siddhartha Jaiswal took an interest. Jaiswal showed that human leukemias, like those in mice, also have elevated levels of CD47 on the leukemic cells. He also found another intriguing link with cancer. When blood-forming stem cells leave the bone marrow to move to another site, they dial up their production of CD47 to protect themselves against macrophages. The way these cells do this looks a lot like the way metastatic cancer cells move around the body and invade tissues. These and other experiments performed by Jaiswal showed that CD47 could indeed play an important role in blood cancers. "But showing that it's possible is different than showing that's what happens in real life," says Jaiswal, who just completed his second year of residency in pathology in Boston.

If leukemia cells can use CD47 to protect themselves against macrophages, then the obvious next question is whether one can reverse that process. Could doctors get macrophages to eat leukemia cells by blocking the CD47 "don't eat me" signal? Majeti, who at that time was still a postdoctoral researcher in the Weissman lab, decided he wanted to take on this project. Majeti identified an anti-CD47 antibody that would block the "don't eat me" signal to the macrophages. He and Chao then mixed labeled leukemia cells with macrophages and the antibody. Under the microscope, they could see the marked leukemia cells inside the immune cells.

'we started out small,

BUT IN THE END

WE WERE GIVING MICE REALLY LARGE DOSES OF ANTI-CD47 ANTIBODY, AND THE MICE WERE JUST FINE.'

In some cases, he could see macrophage cells in the act of eating the cancer cells. Blocking CD47 worked, at least in glassware. After these early tests, the two of them performed the dramatic experiments that showed the antibody worked in mice, too.

When Majeti and his colleagues conducted a full series of experiments with human acute myelogenous leukemia in mice, they were able to totally eliminate the cancers in a majority of the mice. Ash Alizadeh, MD, PhD, another postdoc in the Weissman lab, found that CD47 was also present on non-Hodgkin's lymphoma and performed a similar experiment. He, Majeti and Chao showed that the anti-CD47 antibody, combined with an FDA-approved antibody called rituximab, would eliminate aggressive human non-Hodgkins lymphoma in mice. Rituximab by itself does not.

A MAGIC BULLET?

Killing cancer cells is not that hard. A little household bleach will annihilate the worst cancer doctors have ever encountered. But of course you can't treat cancer with bleach. The trick to all cancer treatments is to harass, inhibit, contain, cut out, beat down and, you hope, kill cancer cells while simultaneously doing as little harm as possible to normal cells in the body. This is especially hard because cancer cells and normal cells are so closely related.

CD47 is not found only on cancer cells. The protein is also on many normal cells, and the obvious danger is that an anti-CD47 antibody would strip away the protective protein cloak from normal cells. Stephen Willingham, PhD, a postdoctoral scholar in the Weissman lab, took on the task of finding out if the use of the antibody would cause macrophage cells to

attack normal tissue. If it did, that would rule out its use as a therapy, no matter how well it eliminated cancer cells.

"The experiments in mice are impressive, but they are rigged in a way because we are using a human form of the antibody against a human cancer, but we are doing it in a mouse," Willingham says. "The humanized antibodies don't attach to the mouse's cells, so the mouse's immune system won't attack its own healthy tissue."

A more fitting test of the safety of the therapy would be to use the mouse version of the anti-CD47 antibody in mice without cancer, which is exactly what Willingham did next. Luckily, these tests showed no major effects on normal tissues in the mouse.

"We started out small, but in the end we were giving mice really large doses of anti-CD47 antibody, and the mice were just fine," says Willingham. The only change was a temporary anemia as the mice's red blood cell count fell (because declining CD47 is a sign of age in red blood cells, blocking CD47 makes the young red blood cells look old to the immune system, which eliminates them). But within days, the level of red blood cells in the mice's bloodstream was back to normal.

"It was actually amazing to me how little effect there was on normal tissue," Willingham says.

CANCER'S 'EAT ME' SIGNAL

How could it be that blocking CD47 is so devastating against cancer but affects normal cells so little? If healthy cells also have CD47, why doesn't blocking CD47 also lead to their destruction? The Stanford researchers hypothesized CD47 can't be the whole story. Cancer cells must also have an "eat me" signal that normal cells do not carry. "It wouldn't be likely that killing cells was the default action of the immune system," Majeti says.

The idea that cancer cells would carry the seeds of their own destruction is not really surprising. Cells have many ways of signaling that not all is well inside them. For instance, specialized proteins inside cells carry bits and pieces of what they find to the cell surface and show it to circulating immune cells. If something is wrong inside the cell, the immune cells can then spot it. It's a bit like a scenario in which parents sit outside their house chatting while the kids play inside. When the kids occasionally come to the window to show them a toy or game they are playing with, the parents know all is well. If a child comes to the window holding a severed human arm, the parents will know something is terribly amiss.

The genetic changes involved in making a cell cancerous disrupt its normal function, making it more likely that the cell will present signs of abnormality, the "eat me" signals that mark it for destruction. Warning signs like these actually make our bodies fairly adept at fighting errant cells. It's likely that every one of us has had cells that are precancerous or cancerous, and that these cells have been effectively dealt with by our body's defenses.

Majeti, Chao and Rachel Weissman-Tsukamoto — a high school student who is also Weissman's daughter — took on the search for an "eat me" signal. They began to focus on a molecule called calreticulin as a possible "eat me" signal because other researchers had shown that it worked together with CD47 to regulate cell suicide. Indeed, the Stanford scientists found calreticulin on a variety of cancers, including several leukemias, non-Hodgkin's lymphoma and bladder, brain and ovarian cancers.

"Our research demonstrates that the reason blocking the CD47 'don't eat me' signal works to kill cancer is that leukemias, lymphomas and many solid tumors also display an 'eat me' signal," says Weissman. "The research also shows that most normal cell populations don't display calreticulin and are therefore not depleted when we expose them to a blocking anti-CD47 antibody."

Understanding how calreticulin and CD47 balance out each other's influence in controlling how the immune system reacts to cancer is important because it can affect how anti-CD47 antibodies are used as a therapy. "If calreticulin is displayed in response to cell damage, you might not want to use anti-CD47 immediately after chemotherapy or radiation," says Majeti. "These treatments can cause damage to normal cells, which might make them vulnerable to macrophage attack when CD47 is blocked."

ONE TREATMENT FOR ALL CANCERS?

One of the many frustrations of cancer is that each type can be so different. New drugs that seem to be effective against one type of cancer turn out to be ineffective against other types. Drugs like Gleevec are miraculous therapies for chronic myelogenous leukemia, but ineffective against acute lymphocytic leukemia. Researchers had high hopes for drugs that block the growth of blood vessels that tumors need, but one such drug, Avastin, while effective against colon cancer, turns out to be ineffective against breast cancer.

As Majeti and his colleagues investigated CD47's role on leukemias, another team in the Weissman lab began to look at solid tumors. Willingham and Jens-Peter Volkmer, MD, a urologist who came from Germany to study stem cells in the lab, began looking at cancer tumor samples collected from patients at the hospital. They were astounded to find CD47 everywhere they looked.

"When Stephen and I first started, we thought we might get lucky and find CD47 on a few tumors, but nobody expected that every kind of cancer we looked at would have

high expression of CD47," Volkmer says. The list of cancers with high CD47 levels ultimately reached 20 and could still be growing. The natural next step was to find out if the methods that were so successful in treating human leukemias could be replicated with solid tumors.

Using collaborative relationships Weissman had built with oncologists at Stanford Hospital over a decade, Volkmer and Willingham obtained biopsies of human cancers, embedded those cancers in laboratory mice and then treated half the mice with the anti-CD47 antibody. As expected, in the untreated control group the samples of deadly, aggressive cancers of the breast, ovaries, colon, bladder, brain, liver and prostate grew rapidly. Special imaging shows how the cancers whipped like wildfire in the mice, starting as one dot of color but growing inexorably over weeks until they had spread throughout their bodies. But in the mice treated with the anti-CD47 antibody, clusters of cancer cells shrank or even disappeared.

"If the tumors were large, then we could shrink them, but if the tumors were small, the antibody could eliminate them altogether," Volkmer says.

Even more dramatic, Willingham and Volkmer, along with pathology professor Matt van de Rijn, MD, PhD, showed that the anti-CD47 antibody could stop cancer from metastasizing, or spreading from the original tumor. This finding is highly important because tumors can most often be cut out or controlled by focused radiation therapy when they restrict themselves to one site in the body. Many cancers become really deadly only once they start spreading the seeds of cancer throughout the body.

"I think the most amazing moment in our research was when we saw that anti-CD47 antibody could stop the metastatic spread of cancers, and even treat cancers that had already metastasized," says Willingham.

As amazing as it seems to the researchers, the CD47-blocking therapy looks as though it could be useful in combating almost every type of cancer (although not every individual cancer — a very small number of cancer samples seem to use other, unknown methods to escape macrophages). It may turn out to be most effective when combined with antibodies that reinforce the positive "eat me" signals from cell surface proteins like calreticulin. "The nice thing is that anti-CD47 antibodies should work with, and boost the effectiveness of, other treatments that use the immune system to fight cancer," says Volkmer.

With the help of a $20 million grant from the voter-funded California Institute for Regenerative Medicine, the Stanford researchers are pushing ahead with plans to begin human clinical trials of the therapy in late 2013 or early 2014. This will be none too soon for many current cancer patients and even the researchers, many of whom spend time treating patients when they are not conducting lab research.

"We are tremendously excited about the potential of CD47 antibodies for improving the lives of our patients," says Beverly Mitchell, MD, PhD, director of the Stanford Cancer Institute.

Majeti remembers many of the patients who graciously donated their cancer cells for research, especially those patients whom he could not cure and who succumbed to their illness. He is sensitive to the cruel irony that in mice, he is able to defeat some of the very same leukemic cells that he could not vanquish in his patients.

"Working in the clinic is a very motivating thing," Majeti says. "When I find myself in a situation that can't be addressed clinically, I say, 'We have *got* to get back to the lab and get on this.'" **SM**

Contact Christopher Vaughan at vaughan1@stanford.edu.

WEB EXTRA A LIST OF CANCERS CARRYING CD47 IS AT HTTP://STAN.MD/KCEFUj

FEATURE

Statistically significant

CONTINUED FROM PAGE 19

Tibshirani is developing methods for analyzing data from phospho-flow cytometry. This technology (pioneered at Stanford) measures protein levels in individual cells, giving a level of resolution that microarrays don't offer. But the data dimensions are flipped: Whereas microarrays compare the expression levels of tens of thousands of genes across maybe 10 to 100 tumors or people, phospho-flow experiments compare the activities of just 50 to 100 proteins across tens of thousands of cells. Thus, new statistical tools are necessary.

"Statistics is a unique field because it's continually reinventing itself based on what types of data are out there," says Nancy Zhang, PhD, assistant professor of statistics at Stanford.

Medicine is now entering a new phase in which "the role of biostatistics will become even more important," Wong says. With the advent of cheap, fast sequencing technology, personal genomes will soon become routine, he predicts. This will open the door for so-called personalized, or individualized, medicine, in which doctors tailor therapies to patients based on their unique genetic makeup.

"The potential of this technology for transforming the way we do medicine is incredible," Euan Ashley says. "And it's really happening." Ashley's team is analyzing the genomes of several heart patients and their families. Identifying their specific genetic defects may help optimize these patients' treatments

(though these are still considered research projects). Cancer doctors are also beginning to sequence their patient's tumors hoping to match patients to the correct therapies.

Sequencing and interpreting genomes brings a host of new statistical challenges. "The technology has moved so fast that we're trying to learn how to leverage it as it arrives," Ashley says. The sequencing machines generate short fragments of DNA reads, containing just 75 to 150 base pairs each, in random order; these have to be assembled into the entire 3 billion base pair genome, which is no easy task. Plus, the technology is imperfect, making an error about once every 100,000 base pairs. "This is a low error rate, but when amplified over the huge amount of data we're dealing with, there are a large number of absolute errors," Dewey says. In the case of the infant girl, Dewey and Ashley singled out a candidate for the causative genetic mutation, but on further investigation saw it was just a sequencing error. The baby continues to have frequent cardiac arrests, triggering the internal defibrillator to restart her heart.

Even if it was possible to reliably make sense of the data, it's still not clear exactly how to use the data to prove that a specific treatment is the best treatment for a particular person. After all, how do you have a clinical trial with just one person? "If I'm going to look at the genome of your tumor and prescribe something uniquely for you, how are we going to assess whether that strategy is going to help or harm you compared with the traditional treatment? There are a whole lot of thorny areas," says Terry Speed, PhD, professor of statistics at the University of California-Berkeley.

Personalized medicine will require innovations in study design. Wong and others are thinking about ways to mine the enormous amount of data stored in electronic medical records (which may soon encompass personal genomes), including the data stored in free text fields. "Physicians are typing away madly. All that information is actually very rich," Wong says.

Biostatisticians are also inventing new ways to do randomized clinical trials. For example, Phil Lavori, PhD, wants to embed randomization into routine care. In situations where the best treatment option is unknown, doctors could enroll their patients (with consent) in an automated clinical trial. "The idea is that a physician could choose option A, choose option B or choose randomize from a drop-down menu," says Lavori, professor and chair of health research and policy at Stanford and a pioneer of point-of-care clinical trials. The institution's electronic system would monitor the trial; and, if a clear winner emerged, would immediately make this alternative the standard of care. A pilot trial of this approach is already under way in Boston comparing two methods of administering insulin to diabetic patients.

Kraemer is proposing new ways of measuring outcomes in clinical trials that better reflect an individual patient's experiences of harms and benefits. For example, rather than comparing the average benefit of a schizophrenia drug with the average weight gain it causes, this balance would be evaluated for each patient individually. Such an outcome measure is more sensitive to individual differences and can help identify which types of patients the treatment is most appropriate for.

Methods for analyzing trial data may also get an overhaul, including allowing knowledge gained in one trial to be incorporated into the analysis of a subsequent trial. "When you take a statistics class in the future, it might not be all about p-values anymore," Lavori says. P-values, which give the probability that a given pattern in the data could have arisen by chance, have been the cornerstone of much of statistics in the past century.

Whatever statistics looks like in the future, one thing is clear: It's a skill set that's only going to get more valuable as biomedicine (and other domains) churn out more and more data. Statisticians can play in any field that interests them, Lavori says. "It's beginning to dawn on people that there are some skills that stay valuable no matter what happens, and I would say that statistics is one of them." **SM**

Contact Kristin Sainani at kcobb@stanford.edu

Executive Editor:
PAUL COSTELLO
Editor:
ROSANNE SPECTOR
Art/Design Direction:
DAVID ARMARIO DESIGN
Director of Print and Web Communication:
SUSAN IPAKTCHIAN
Staff Writers:
MICHELLE L. BRANDT
KRISTA CONGER
ERIN DIGITALE
BRUCE GOLDMAN
KRIS NEWBY
RUTHANN RICHTER
KRISTIN SAINANI
JOHN SANFORD
CHRISTOPHER VAUGHAN
TRACIE WHITE
SARA WYKES
Copy Editor:
MANDY ERICKSON
Circulation Manager:
SUSAN LYDICK

Stanford Medicine is published three times a year by the Stanford University School of Medicine Office of Communication & Public Affairs as part of an ongoing program of public information and education.
 Letters to the editor, subscriptions, address changes and correspondence for permission to copy or reprint should be addressed to *Stanford Medicine*, Stanford University School of Medicine, 555 Middlefield Road, Building 110, Menlo Park, CA 94025. We can be reached by phone at (650) 736-0297, by fax at (650) 723-7172 and by e-mail at medmag@stanford.edu.

To read the online version of **Stanford Medicine** and to get more news about **Stanford University School of Medicine** visit http://med.stanford.edu.

For information from the **Stanford Medical Alumni Association** visit http://med.stanford.edu/alumni/.

COMFY MICE

IMPROVING LIFE IN THE LAB

Joseph Garner's work easing life for laboratory mice took a turn in 2004 with an incidental observation at a scientific talk. The speaker — a colleague and future collaborator — presented a graph showing that mice, who prefer balmy climes in the upper 80s, experienced increased metabolic rates in cooler laboratory temperatures. • "What was so striking about the graph," recalls Garner, PhD, an associate professor of comparative medicine who's studied animal welfare for more nearly two decades, "was how stressed the mice would be by the temperatures we normally house them at." • He began to think that the chilly temperatures, mandated by federal rules, could help explain why one out of 10 drugs successfully tested in mice end up not working in people.

"If you want to design a drug that will help a patient in the hospital, you cannot reasonably do that in animals that are cold-stressed and are compensating with an elevated metabolic rate," he says. "This will change all aspects of their physiology — such as how fast the liver breaks down a drug — which can't help but increase the chance that a drug will behave differently in mice and in humans."

For a mouse, life in a nest is best. Researchers found that lab mice given nesting material use it to create cozy homes.

Garner is among a handful of researchers studying the issue and advocating for mice, hundreds of millions of which live in research labs. The work already has led to changes in animal care regulations and spurred more interest in the psychological, as well as physical, well-being of the animals, he says.

Garner is a great admirer of mice, who are among the planet's most flexible of creatures, able to live alongside humans in even the most uninhabitable places.

"Our shared history makes them the perfect research animal," he says.

Laboratory mice are kept in the cold because it suppresses their aggressive tendencies. Raising the temperature would make them unmanageable.

So Garner and his colleagues found a simple solution. In a recent study, they observed that if mice are supplied with shredded paper, they will build a cozy nest that allows them to naturally regulate their temperatures to a comfortable level.

The mice could move between cages of varying temperatures and with varying amounts of nesting material, telling the researchers how nesting material compensates for cold temperatures. The animals routinely chose a warmer locale, though some of them wanted to have it all — a warm spot and a nice little home, too.

"They would go on holiday somewhere AND take their nest with them," Garner says. "Some people like to take a pillow on holiday and some don't. These mice were packing their own pillow."

He says nests also help decrease the animals' stress and anxiety.

"The really obvious explanation is that they can hide in it — hide from us, because we are their predators. Imagine you are hiding from a Tyrannosaurus rex — you want a couch to hide under. That is a little like what's going on," he says.

He says mouse nests can help researchers in other ways as well. "The shape of the nest tells an experienced person whether the animals are too hot or too cold, whether they are sick or whether they are about to give birth," Garner says. "Once you learn how to 'speak mouse nest,' the nest is a wonderful tool that anyone can use to assess the general state of the mouse." — RUTHANN RICHTER

BRIANNA GASKILL

S T A N F O R D

M E D I C I N E

Stanford University School of Medicine
Office of Communication & Public Affairs
555 Middlefield Road, Building 110
Menlo Park, CA 94025

Change Service Requested

Blond roots

PACIFIC ISLANDERS' GOLDEN-LOCKS MYSTERY IS SOLVED

BLOND HAIR IS RARE. SO GENETICIST SEAN MYLES was amazed by what he saw when visiting the Solomon Islands, an equatorial nation in the South Pacific. • "They have this very dark skin and bright blond hair. It was mind-blowing," says Myles, who later became a postdoctoral researcher at Stanford, working with genetics professor Carlos D. Bustamante, PhD. • "As a geneticist on the beach watching the kids playing, you count up the frequency of kids with blond hair, and say, 'Wow, it's 5 to 10 percent.'"

Myles later learned that blond hair occurs with substantial frequency only in northern Europe and in Oceania, which includes the Solomon Islands and its neighbors. He was so intrigued he arranged a return trip in 2009 to figure out the genetics of the blond trait. He and a friend, scientist Nicholas Timpson, PhD, gathered the data — including hair color measurements and DNA (from saliva samples) — and then they worked with Bustamante and others to analyze it. They published their discovery May 4 in the journal *Science*.

Many had assumed the islanders' blond hair was the result of gene flow — a trait passed on by European explorers, traders and others who visited in the preceding centuries. The islanders themselves gave several possible explanations for its presence, says Myles, who is now an assistant professor at the Nova Scotia Agricultural College. They generally chalked it up to sun exposure, or a diet rich in fish, he says.

But this new study shows that blond hair among indigenous Solomon Islanders is the result of a homegrown genetic variant that's distinct from the gene that leads to blond hair in Europeans. So blond hair arose in both places independently.

Aside from the "gee whiz" factor, the discovery underscores the importance of genetic studies on isolated populations, says Bustamante. "If we're going to be designing the next generation of medical treatments using genetic information and we don't have a really broad spectrum of populations included, you could disproportionately benefit some populations and harm others."

Bustamante is seeking funds to analyze the rest of the data gathered. "For instance, the genetics of skin pigmentation might be different there too — not the same as in Europe or Africa or India. We just don't know." — ROSANNE SPECTOR

CONTACT *STANFORD MEDICINE*
at medmag@stanford.edu
or call (650) 736-0297.
Follow @StanMedMag on Twitter.

SEAN MYLES