CHALLENGES AND OPPORTUNITIES:
GENOMIC DATA, PATIENT CARE, AND THE CLOUD

PROVIDED BY

healthcare informatics
CUSTOM MEDIA

IN COLLABORATION WITH

NetApp®

Introduction

Patient care organizations—primarily academic medical centers affiliated with research universities that are conducting clinical research into all forms of cancer and into other diseases with genetic components—are moving forward to connect the academic research arms of their universities to the patient care delivery operations in their clinical organizations. And that is leading both to opportunities and challenges.

On the opportunity side, genomic data is now actively being used for rare disease diagnosis; for cancer detection; for the tracking of mutations; and for medication selection for patients.

But the data challenges involved in working with genomic data, particularly in participating in any activities connecting genomics to patient care, are many, and complex. Here's how Dr. Hakon Gudbjartsson, Ph.D., the CIO of WuXi NextCODE, a genomic information company and global platform for genomic big data, put it in a recent webinar: "If you're thinking that genomics isn't a big data challenge, think again. Although all humans share the same DNA, the sequence of it is 100-percent unique to the individual. If you and a friend were to compare your DNA, you would find about 5 million differences." Over the past 20 years, the leaders at WuXi NextCODE have amassed the world's largest database of human genome sequences. "The challenge is to take a dataset of 5 million and figure out the differences or mutations that are important — which ones are the causes of rare diseases, which ones are the causes of cancer, and how to treat patients," Dr. Gudbjartsson explained.

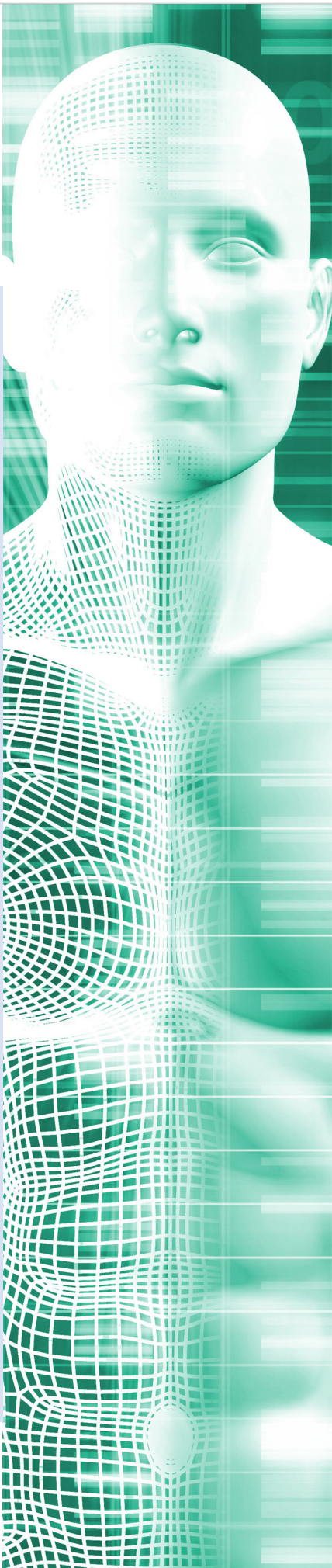
And, even as the potential exists in medicine to facilitate better diagnoses, earlier interventions, more-efficient drug therapies, and customized treatment plans, in what is being called, variously, personalized medicine, precision medicine, and individualized medicine—which provides a genomic blueprint to determine each person's unique disease susceptibility, define preventive measures and enable targeted therapies to promote wellness—the challenges are also legion.

"Genomic research is going very well, and we have thousands and thousands of academic papers around it. But how you translate the research into patient care is complicated, and we're at the beginning stages," says Sainand Balu, Facility Director in the Bioinformatics Core at the Lineberger Comprehensive Cancer Center at the University of North Carolina, Chapel Hill, in Chapel Hill, North Carolina. "And how the successful push to get electronic health records [EHRs] implemented in hospitals, dovetails with the storage and sharing of genomic data, is a complex story. Genomic data is messy, and its interpretation is both important and challenging."

Indeed, by 2019, volume of data generated by genomic sequencing will be greater than the amount of data on YouTube.

Among the data management challenges that industry leaders have articulated are the following:

- The size of files: 100 gigabytes each, when it comes to whole-genome sequencing.
- The volume of files: new sequencers can generate in excess of 1 terabyte per day; and the total volume of genomic data more than doubles every year. It is estimated that it will exceed 6.2 exabytes by 2019.
- The growth of Federal genomics programs and data repositories.
    - The NIH recently introduced the STRIDES program which will help organizations to increase the sharing of genomic data. Google was selected to assist with the first phases of the program.
    - Another new NIH program is called "All-of-US." The objective of the program is to collect DNA samples and medical history information from a million volunteers. This will give genomic researchers a large and diverse data set to analyze.
- The cost of DNA testing is dropping, particularly through the commercialization of consumer-available testing. More and more consumers are shipping a DNA sample to learn more about their family history and to possibly identify genomic mutations which indicate a higher risk of a diseases like breast cancer. Some healthcare professionals are concerned about the accuracy of the consumer driven DNA test kits. They argue that the DNA test data is not reliable enough yet to do a complete diagnosis of a disease. Consumers need to get a second opinion from their primary care doctor. However, the low cost and convenience of consumer DNA test kits indicates that the trend will continue. And, a 'tidal wave' of genomic data will continue to be a data management challenge for many healthcare and research institutions.
- Government regulations around data security and privacy may complicate how genomic data is generated, used, shared, and stored. Historically, genomic researchers simply de-identified the genomic files by stripping out the DNA donor's name and other personal information. While the practice of de-identifying data has been acceptable for years to the overall genomic research community, there are serious security and privacy questions on the horizon.
    - For example, we know that an important goal of personalized medicine is to accelerate the sharing of an individual's genetic profile information with clinicians—at the point of care. This means that HIPAA rules and guidelines for safeguarding protected health information, PHI, may apply to the patient's genetic data as well. In cases where a patient is participating in a clinical trial for a new medicine or treatment, this research process may require the de-identifying and re-identifying the individual throughout the trial.
    - Another potential data management challenge for genomic researchers and clinicians is the new General Data Protection Regulation, GDPR. This EMEA focused program became effective on May 25th, 2018. And, its impact is global. For example, a data scientist from Germany may visit the U.S. to conduct genomic testing with her or his colleagues. Interestingly, the GDPR regulations may apply to the data scientist. And, if she/he is hospitalized during the visit to the U.S., then the GDPR rules will potentially apply as well.



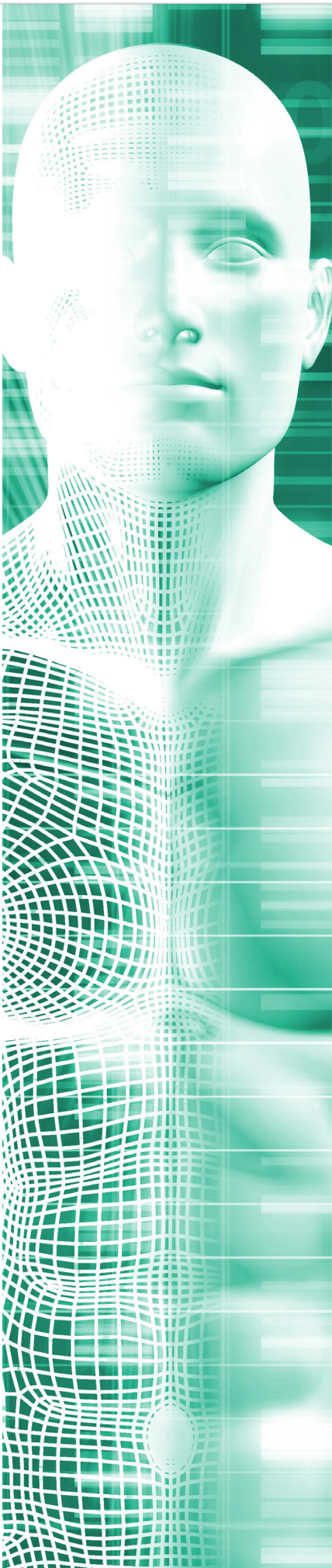
What About the Cloud?

For information technology and clinical informatics leaders in healthcare, one of the biggest practical questions is around how to manage, share, and store the massive amounts of data and sizes of data loads involved in genomic medicine.

As Ben Langmead and Abhinav Nellore noted in an article published online on January 30, 2018 in Nature, and entitled "Cloud computing for genomic data analysis and collaboration," "The cloud's chief advantages are elasticity and convenience. Elasticity refers to the ability to rent and pay for the exact resources needed, and convenience refers to the fact that the user need not deal with the disadvantages of owning or maintaining the resources." Still, they noted, "Archives of sequencing data are vast and rapidly growing. Cloud computing is an important enabler for recent efforts to reanalyze large cross-sections of archived sequencing data." As the authors note, "The cloud is becoming a popular venue for hosting large international collaborations, which benefit from the ability to hold data securely in a single location and proximate to the computational infrastructure that will be used to analyze it." What's more, they note, "Cloud clusters can be configured with security measures needed to adhere to privacy standards, such as those from the Database of Genotypes and Phenotypes (dbGaP)."

But the challenges involved are immense. In an article published in the November 2015 issue of The Journal of AHIMA bluntly entitled "Can EHRs Handle Genomic Data?" Daniel DuBravec, CHTS, CEHRS, a senior consultant at LMI, a non-profit government consulting firm, who consults for the government on EHR patient privacy and security standards, wrote, "In an article on the technical needs for integration of genomic data into EHRs published in the Journal of Biomedical Informatics, the authors express valid concerns about the enormity of raw genomic data: 'The large volume of each individual's DNA, protein and related data—hundreds of gigabytes to terabytes in its raw form—exceeds the capacity of commonly available network bandwidth and disk storage in healthcare settings.' Geneticists are predicting that the storage needs for patient genomic data at the capacity of 2-40 exabytes of storage, exceeding the overall data storage used by YouTube, as an estimated 100 million to two billion human genomes are sequenced by 2025. Rapid retrieval and analysis of this data is challenging for the typical database system used by most EHRs, leading to substantial performance issues," he added.

DuBravec also argued for extra-EHR storage, writing that, "By storing the raw data outside of the EHR, users can transfer data sets into the system when they need them. The original raw data could be stored locally in the hospital in storage clusters or remotely in 'the cloud.' Cloud-based storage and applications companies such as Amazon and Google are currently developing the infrastructure to host the raw genomic data to alleviate the technical and infrastructure burdens from hospitals. Security concerns over where the data is stored, whether in the cloud or in different clusters, remain and must be compliant with federal Health Insurance Portability and Accountability Act (HIPAA) Privacy and Security Rules."



DuBravec sees ongoing challenges in terms of the searchability of EHRs for readily findable genomic data. His article continues, "Thanks to browsing tabs, drop-down menus, and search fields, traditional data such as patient vitals and medical history are becoming more searchable in most EHRs. As clinicians seek pertinent information, the navigational design of the EHR provides search capabilities to easily find patient data. It even has the ability to create impromptu reports on the patient's condition. Most EHRs attempting to include genetic information provide a clinical notes section where free form text can be captured. This data is either automatically imported from laboratory systems or copied and pasted, where it lives undiscoverable and difficult to interpret. Another method used by healthcare providers is to attach a PDF as part of the patient record." Clearly, such workarounds are unsustainable on a long-term basis and at scale.

"Thanks to browsing tabs, drop-down menus, and search fields, traditional data such as patient vitals and medical history are becoming more searchable in most EHRs."

— *Daniel BuBravec*

Dr. Hakon Gudbjartsson, CIO of WuXi NextCODE, believes there are additional alternatives. "Another consideration is that many of the genomic research organizations will eventually utilize a 'hybrid' cloud infrastructure model which gives them the best of all worlds," Dr. Gudbjartsson states. "Specifically, we expect research institutions to utilize both "on-premise" infrastructure like CAPEX, as well as the 'off-site' cloud services models like OPEX.

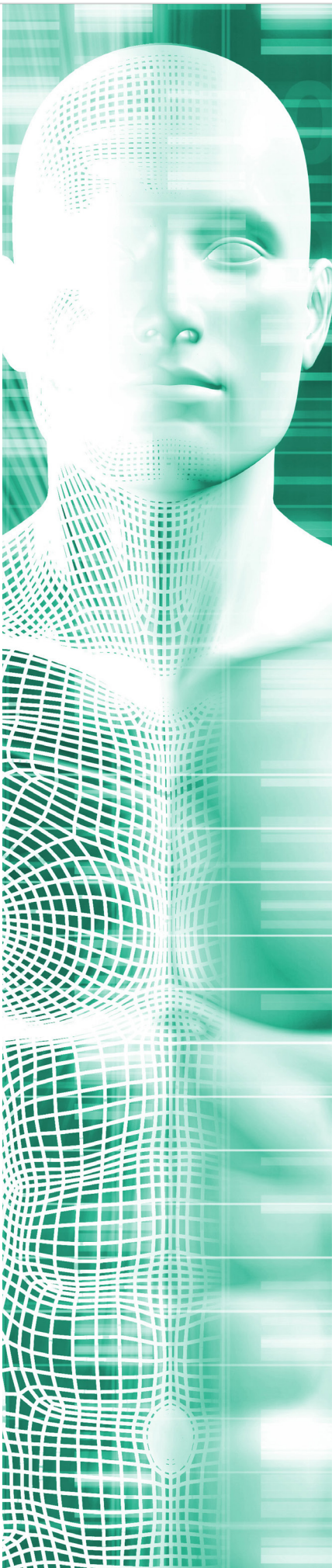
What Healthcare IT Leaders Have to Say

So, what do healthcare leaders in the trenches with this work have to say? All agree that it remains challenging; though some points of light are beginning to emerge. Karl Poterack, M.D., Medical Director of Applied Clinical Informatics at the Mayo Clinic, sees the overall landscape, and Mayo's advancements within that landscape, clearly. Asked what the biggest challenges are right now, Dr. Poterack says, "I think that there are three or four big challenges overall. One is the sheer volume of data that is involved. There's a huge volume of raw data when you start talking about genomics; and there's a huge volume of, what does this gene do, what might it do, what is it associated with, how strong is the association? There's a huge volume of that, and very few people are experts in all of it; you need some sort of guidance. A lot of information is out there, including where there's not expertise."

Meanwhile, Dr. Poterack says, "The other big issue is that you get a pile of data, and there are pieces of data that are erroneous or inaccurate or misleading. You put somebody through a test process and get a huge pile of genomic data, and very similarly to if you do a whole battery of lab tests on somebody, if you do enough lab tests, some will be abnormal, and some will be erroneous or won't mean anything. And because it's genomic, everybody is expecting a lot from this, and understandably so, but in the end, it's just more pieces of data, and when you have a lot of data without context, you can be really misled."

A further challenge? While a disconnect exists between the academic research and clinical sides of academic medical organizations along numerous dimensions, Dr. Poterack notes that, "We see that all the time, but it's particularly pronounced in this area, for a number of different reasons. First of all," he says, "the average physician just doesn't have the depth of understanding of genomics as an academic topic, to really be able to able to communicate in that world; at the same time, the more purely academic folks who aren't clinical, don't fully understand, since they don't live in the clinical world, how messy the processes around clinical medicine can be." In that context, he says, asked what the best practices are in this area at the moment, "The best practices are making expertise available to clinicians. In my own institution and others as well, one of the areas on the leading edge of this is clinical pharmacy departments that are developing expertise

> "Because [this data] is genomic, everybody is expecting a lot from this, and understandably so, but in the end, it's just more pieces of data, and when you have a lot of data without context, you can be really misled."
>
> — *Dr. Karl Poterack*

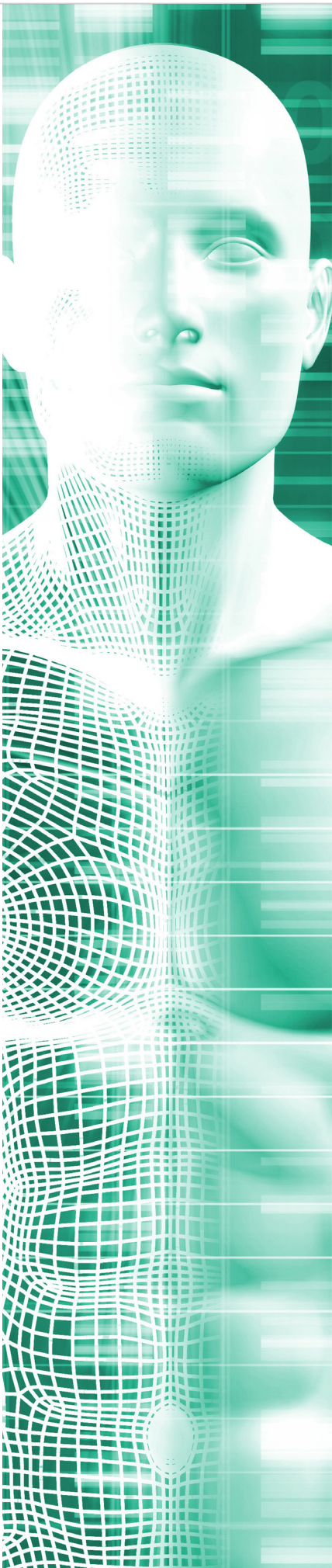


in pharmacogenomics, and can say, OK, if this is the issue you're concerned about, here are the tests you can run, and once they've been run, they can say, OK, this is what those results mean. Whether it's the clinical pharmacists sitting down with the patients, or going over the results with the ordering clinicians, and saying, this is what those tests mean. Without that, the clinician just has little hope of navigating this."

And that is exactly what is taking place these days at Mayo Clinic, Dr. Poterack notes. Speaking as a still-practicing anesthesiologist, he says, "One thing that we anesthesiologists have all seen over the years is that people come for a surgical procedure—and patients will say, this pain medicine doesn't work, I prefer this other one. Or, this pain medicine has horrible side effects. And we've heard from these patients, and sometimes, what they say is so out there and they're so vocal about it, so that you almost wonder whether there's something psychological going on. Well, come to find out, we've got folks doing clinical pharmacogenomics who now have assays and can do genetic testing. And it turns out we can have very interesting combinations of genetics in people so that a particular pain medicine actually will not provide them much analgesia, but will make them sleepy and dopey—but keep them in pain. So we didn't have a good pharmacological explanation, though maybe they were drug-seeking or something. Now, when we have these people with these odd reactions to pain meds, we can send them to these clinical pharmacists, they can run tests and run reports, and can say, here's their genomics, here are the pain meds that work for them. So, they have specific genomic markers that make them poor candidates for specific pain meds."

Importantly, Dr. Poterack notes, at Mayo Clinic, "Thankfully, the raw data is not being just dumped into the EHR, if for no reason that it's a tremendous amount of data. So even if somebody figured out how to make it show up, it would take that much more effort to make it show up meaningfully. So that's not happening, and that's good. A lot of this is very purpose-built" at Mayo, "we've got a battery of assays we can do to sort of help characterize somebody's genes with regard to how they metabolize certain drugs—we, being a clinical pharmacy unit, a specific genomics unit, laboratory services, or whoever's managing this. When it's being done usefully, it's being presented as a report. Rather than getting a bunch of lab tests that say your sodium is this and your potassium is that, and by the way, this gene is heterozygous for such and such, it's being presented with meaning, saying, these particular medications won't metabolize. It's being presented in a report that walks the clinician through the implications of the results. Those are the really effective ways this is being done." Indeed, in effect, the way in which such data and information are shared with physicians caring for patients is akin to receiving a specialty consult, he adds.

Fortunately, the landscape around the storage and sharing of genomic data in the context of patient care delivery is beginning to shift now, says David LaBrosse, Director of Emerging Healthcare & Life Sciences Solutions at NetApp. "We can provide on-premise and cloud-based IT solutions which address many of the genomic data management challenges, today."



While electronic health records have been well established, integrating the genetic data is not advancing as quickly as it should. LaBrosse says this is an issue because clinicians do not have easy access to genetic test results. And, this data gap can impact patient care. For example, when a patient is battling breast cancer he or she may need genetic tests to determine if there is a family history of the disease. If the test results show mutations in BRCA1 or BRCA2 genes, then he (and his extended family members), may have a higher risk of breast cancer. That is important data which should be stored in electronic health records. The good news is that healthcare institutions are beginning to address the genetic data gap.

LaBrosse sees multiple phases of genomic data management, from collection through sequencing, through analysis, to the application of insights to direct patient care.

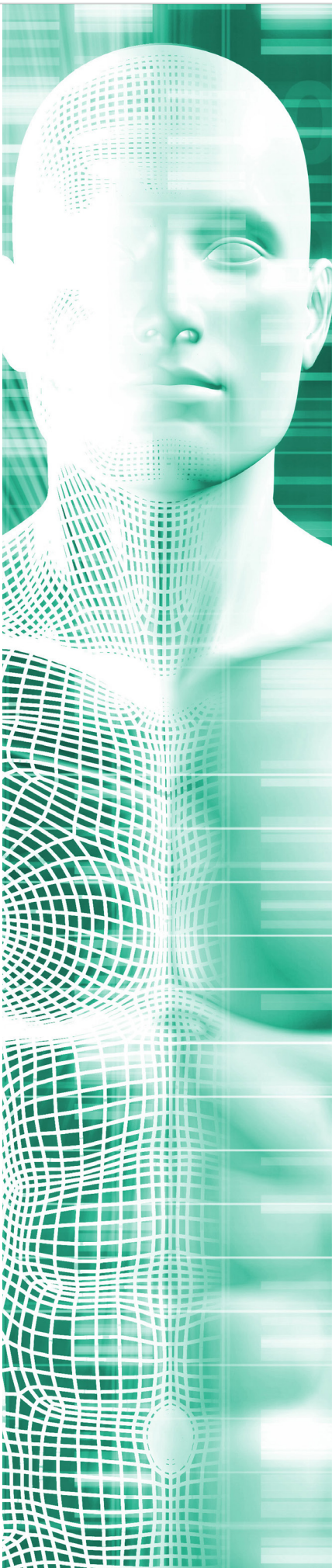
The collection phase has changed in recent years. While many research centers rely on hospital-based bio-banks for DNA samples, some genomic research organizations are signing contracts with consumer DNA kit vendors—like 23andME. In fact, DNA kit vendors have collected millions in DNA samples in only a few years. This spike in recreational genomic testing is creating challenges for clinicians. Patients are asking hospitals to add the DNA test results into their electronic health records. Verifying the accuracy of the genetic tests may become an issue.

It is in the sequencing phase where the size of the genomic files can become a major issue, LaBrosse notes. "One file involving whole genome sequencing can be as large as 100 gigabytes. And, the complete human genome sequence process may include meta data which brings the file size closer to 200 gigabytes. That's huge. And that's where the initial IT headache begins, because you've got such large files to manage and protect. Plus, depending on the size of the lab or data center, they may or may not have the capabilities to scale." These research centers are basically facing a tidal wave of genomic files.

"The complete human genome sequence process may include meta data which brings the file size closer to 200 gigabytes. That's huge. And that's where the initial IT headache begins."

— *David LaBrosse*

The reality, LaBrosse says, is that "There are data centers sitting out there with large silos of genomic data. Some are not able to share their data because of the costs of migrating the files on-premise or in clouds. Backing-up the genomic files is a headache as well. Many research centers are running out of space. And, power and cooling expenses are rising.



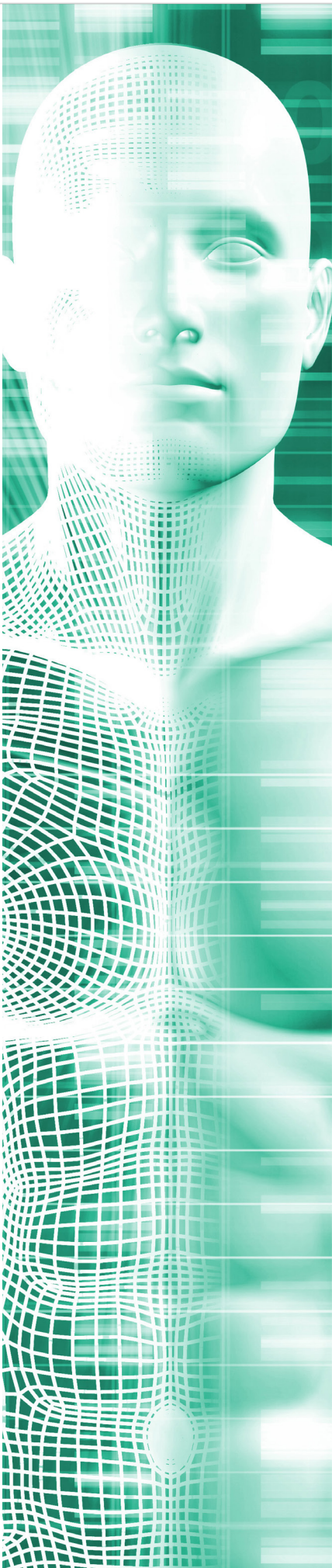
So, IT people have to figure out ways to manage this massive amount of file structure." Inevitably, he says, that means "some level of file compression. There's a software company in the U.K. called PetaGene that can take a large genomic file and compress it down to around 30 gigabytes. We have conducted tests with the PetaGene software. And, we validated their file compression capabilities. It is impressive. They are capable of going into large research institutions and saying, why don't you give yourself some breathing room?

Meanwhile, during the analysis phase, LaBrosse says, there are multiple complexities involved. "Genomic files initially go through quality assurance steps to ensure the sequencing process is accurate. This sometimes requires high performance analytics. Once the genomic file is of use, they go into the additional analytics phases, applying homegrown or industry standard algorithms to discover new mutations. Some research teams create their own algorithms, because so much is still not understood in genomics today. I think the genomic industry is striving to get to a more standardized approach. However, with millions of DNA differences to analyze, there will be plenty of genomic analysis ahead.

During the later stages of genomic analytics, the focus is on providing useful genetic information for doctors, oncologists, geneticists and other clinicians. LaBrosse says. "Genomic researchers are brilliant and do amazing things; but we still need to expedite the delivery of genetic test results. We need to send meaningful data to the point of care as rapidly as possible. At the end of the day, we need to remember that the final stages of genetic testing are applying what we've discovered, to help patients and their families to improve their health.

Some of those working in patient care organizations see hope in certain forms of collaboration. "Standardization, presentation, and interoperability, are all big challenges," says the University of North Carolina's Balu. "These are really large data sets. How can the treating physician interpret the data so that he or she can make decisions based on it? Even bringing in the genetic test markers—how you make the data usable for the physicians, is a huge problem. Over the last few years, one trend on the clinical care side is the effort to build learning healthcare systems. They want clinical decision support systems that can continuously learn data fed into it. For a learning system to work, you need clear data standardization. So how do you make sure that the data is presented correctly in your EHR so that it can be used in your research, when you're collaborating over the care of hundreds, or even thousands, of patients?"

Balu says that part of the answer will inevitably involve collaboration around standardization of data and of data presentation. "There's a consortium, the Global Alliance for Genomics and Health, a large consortium that is trying to bring about standards in representing genomic data in a healthcare context," he reports. "It's moving forward. But the problem is that the genomics area is such a rapidly changing area. So what we think of as a gold standard two years ago, is not a gold standard anymore. Also, there are more and more tools that come up. And with newer tools, you're extracting more data. So how do you view data that's come from old data sets? How do you make sure that data is standardized and presented in a reasonable fashion, to move forward both for research and patient care?" All of those questions remain to be answered, as genomic data increasingly enters the realm of patient care delivery.



What Should Healthcare IT Leaders Do?

What should CIOs, CTOs, CMIOs, and other healthcare IT leaders "I think that, speaking from my personal experience, the technology aspects of it—most CIOs, CTOs, etc., in the healthcare environment, are more focused on clinical care," says Balu. "They have a difficult time understanding in a clinical care environment, how these large-scale computing environments should work—how you do research and bring it back into patient care delivery. So organizations need to make the strategic decision that research analytics is an important arm of a healthcare organization in itself. And there are practical technological decisions to be made—building infrastructure, creating data standards, supporting privacy and confidentiality for patients, and how you share the data. Data use agreements need to be worked on at some level. But to me, a strategic decision has to be made that this is important, and a commitment has to be made."

What's more, Balu says, "The technical problem is more around how you make sure that huge amounts of data are stored, shared, and analyzed. You have to validate data for patient care. At the same time, hospitals and hospital IT environments have a very difficult time—most hospitals don't even yet have high-performance computing (HPC)—they may have sequencing machines, but they run minimally needed analysis, and it ends there. There are a lot of researchers in the hospital environment who want to run large-scale research, and need data to be shared without too much burden; that would be a great improvement."

"The CIO needs to do an assessment of where genomic data fits and look at the holistic picture of the genetic data coming their way," NetApp's LaBrosse says. "The pressure isn't slowing down, and CIOs and other IT leaders have to be ready."

Looking forward into the next decade, the challenges and opportunities both loom large. For the leaders of patient care organizations, there is tremendous opportunity in bringing genomic research and data directly into the patient care setting. But the range of policy, process, and technological issues remain daunting. And healthcare IT leaders will need to make wise decisions around strategic planning for these efforts, and around the technological investments they will need to make, as their overall investment in this area grows and accelerates.

> "Organizations need to make the strategic decision that research analytics is an important arm of a healthcare organization in itself. And there are practical technological decisions to be made."
>
> — *Sainand Balu*

Key Takeaways

- Genomic research is bringing increasing amounts of valuable data into the patient care delivery sphere, primarily in academic medical environments, but potentially beyond academic environments as well.
- The data infrastructures of most patient care organizations have not yet been organized to ingest genomic data and information, at least not at any level of scale.
- The leaders of a growing number of patient care organizations are developing early strategies around how to ingest and use genomic data in an organized way and bring it to physicians in practice in ways that are effective and useful.
- Cloud computing will be one key mechanism for helping to manage the very large file sets of data involved in ingesting genomic data into electronic health records and clinical information systems.
- Early collaborations are beginning to point the way to standards for data organization and presentation.