HOW YOUR BUSINESS WORKS
information management
February 24, 2014

Integrating OMIC Data with Patient Records to Improve Clinical Outcomes – a Big Data Approach

By Todd Saunders

Before the Affordable Care Act (Obamacare) became the law of the land, proponents marketed it to skeptics as an amalgamation of Republican and Democratic proposals in one package. The most widely used example was Massachusetts' state-wide health care initiative implemented under a Republican governor. One fact that gets overlooked, however, is that it was actually George W. Bush's administration that first promoted a national mandate for an electronic health record system. "In 2004, President Bush set as a goal that every American would have an electronic health record by 2014." 1In 2014, it's now standard practice for a physician to greet his patients with a handshake and a computer tablet. By encouraging physicians to create EHRs for each of their patients, Bush's administration set in motion a practice that would, 10 years later, prove invaluable in providing some of the big data necessary toward understanding and treating complex diseases, developing more effective, individualized patient treatments, and reducing costs.

EHRs are an important part of a much larger puzzle. It's now widely accepted that big data can significantly transform health care delivery and improve outcomes for patients everywhere. The "Policy Forum on the Use of Big Data in Health Care" states: "Big data has the power to transform lives. In health care it can reveal the factors that influence health, help target appropriate care for individuals or populations, enable new discoveries, shape outcomes, and reduce costs."2

Additionally, the authors of the report, "Embracing the Complexity Of Genomic Data For Personalized Medicine," point out that "Numerous recent studies have demonstrated the use of genomic data, particularly gene expression signatures, as clinical prognostic factors in cancer and other complex diseases."3 Big data, in this context, refers to all of the types and varieties of data that are now available and can be integrated with patient information (both health care related and non-health care related) to provide more targeted

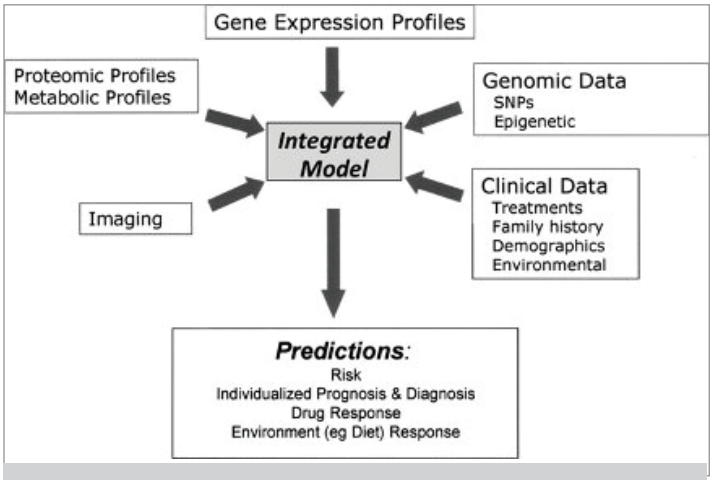

Figure 1: An integrated use of genomic, clinical and other data to predict clinical and biological phenotypes7

and effective care and treatment, also known as personalized medicine. Among the data now available is data that is referred to as omic data. In health care, the more popular types of omic data currently being studied include genomic, pharmacogenomic, and proteomic (just to name a few):

- **Genomics** is the study of the genome (both the genes and the non-coding sequences of the DNA/RNA) of an organism.4
- **Pharmacogenomics** is the study of the impact of specific pharmaceuticals relative to a person's genome.5
- **Proteomics** is the study of the entire set of proteins produced by an organism.6

Data scientists have already accomplished considerable research in these fields. Rich sets of this type of information can be accessed and are often available for download from institutions such as the CDC, NIH, NCI, FDA, HHS, CMS, and many academic institutions. When this data is combined with patient-specific information, diagnoses and treatments have the potential to be more accurate and more effective, thus improving outcomes.

By including all this data, we can provide new insight into the whole health care ecosystem. We also exponentially

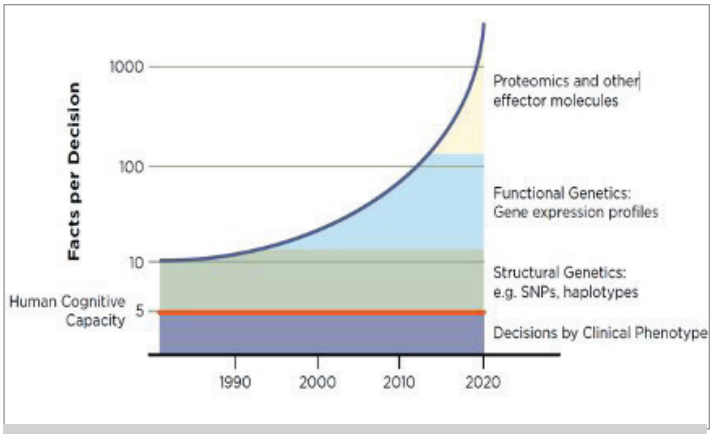

Figure 2: Medical decisions becoming more complex

increase the amount and complexity of information available for decision-making. The authors of the genomic data for personalized medicine report explain state: "Clinical disease states represent exceedingly complex biological phenotypes reflecting the interaction of a myriad of genetic and environmental contributions."8 No matter how smart some of us think we are, our human ability to process this amount of information doesn't hold a candle to the current capabilities of big data's processing power. The following chart from "The Learning Health System and its Innovation Collaboratives Update Report" illustrates how much information a typical human is able to leverage in their decision-making process versus how much information is actually available for decision-making, specifically in the case of diagnoses made by physicians, in the coming years.9

Note that the scale on this chart is a log scale, and the amount of information begins growing by orders of magnitude. This indicates that the amount of information available to assist with clinical decision-making has already greatly overwhelmed a human's ability to process it.

Practical Applications and Solutions Available Now

This is where big data solutions in clinical settings enter the picture. Humans (e.g., physicians and caregivers) need help to effectively analyze the available data and use the results of that analysis to determine the best courses of action relative to current circumstances. Exciting advancements in big data and data science allow us to configure a big data solution to perform much of the required analysis (e.g., genomic pattern matching) necessary to integrate clinical and research data with a patient's unique EHR, and then provide only the relevant and salient results to a caregiver. This pinpoint process allows the caregiver to make decisions based on a manageable amount of highly filtered and relevant information. For example, certain cancer drugs have far less efficacy if a certain genotype is present in a patient. Now that we can analyze the patient's genome prior to prescribing a cancer treatment, a more effective course of treatment can be prescribed immediately rather than waiting weeks or months, only to determine the first prescribed drug was ineffective.

We've learned from our work in this area that the key to bringing together and integrating all of these types of data is a very flexible, accommodating solution architecture. The solution architecture provides a way to link to and/or accept data into it. It also integrates and models the disparate types of data into structures that make sense for the desired usage of the integrated data. Finally, it provides an appropriate interface to facilitate the needed functionality such as reporting, informatics, modeling, research, and eventually, predictive diagnoses. From a technical perspective, the solution architecture must be both scalable (can handle bigger volumes of more types of data and support more users) and extensible (can perform enhanced and new functions).

Because the types of data are varied (e.g., structured data, lab results, text files, scanned images, scanned text, unstructured data, etc.), the solution requires more than one type of data store. The most common type of data store over the last couple decades has been relational databases. This type of storage is very effective for managing structured and well-understood data in support of data integration as well as reporting and analytics. However, it's not as strong at managing extremely large volumes of data and less-structured data such as raw text or documents.

The need to accommodate unstructured data led to the development of different types of data stores, the most popular currently being Hadoop, though other good solutions for unstructured data also exist. This type of data store is best for storing and archiving data as well as doing batch loads of large volumes of data. Its redundant architecture protects data and reduces the need for system backups.

Another common type of data store is columnar databases. They are similar to relational databases, and can often be queried using identical tools, but they store data in such a way that greatly improves response times on queries performing analysis on large volumes of data.

By combining these three types of data stores into a common platform, nearly every type of data and analysis can be supported. Like most big data solutions, the key to the success of this platform still resides with the actual data management protocols, using a robust metadata approach. For data ingestion, all data must comply with the defined ontologies, taxonomies, business definitions and syntactic standards. The loading and integration of the data is

effectively managed with rules supplied within the data integration metadata, and the access to and use of the data is controlled through a comprehensive semantic layer. These mechanisms help ensure the end user is getting the data they intend to get and using it appropriately, according to their level of data access as defined by privacy laws.

Many facets of this overall solution are already well-established. Health Information Exchange software solutions allow for the integration of EHRs and other data across many different facilities and organizations. These software packages have built-in capabilities to connect to the all the leading EHR packages and the logic to consolidate patient records and other information. Hadoop and other NoSQL solutions are widely used to analyze the vast amounts of omic data that is produced. Newer and more powerful statistical methods are constantly being developed to improve the understanding and use of the omic data.

In the end, it boils down to discovering a transformative quality of care using a patient's EHR data with research data and omic data to get down to the individual patient to determine variants and isolated factors relating to a person's specific DNA/genomic identity. When this omic data is combined with patient EHR data in real time, and caregivers can access a wide variety of research data/clinical test data at the same time, it greatly increases the efficacy of treatment for patients, minimizes the need for re-visits and reduces costs. The good news is that the technology and big data resources are already in place toward achieving an underlying architecture that can support the wide variety and high volumes of data and make it accessible to researchers, clinicians and physicians in a manner that supports their decision-making processes.

References:

- March-April 2008, Journal of the American Medical Informatics Association, "Promoting Electronic Health Record Adoption. Is It the Right Focus?" Donald W. Simborg, MD;. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2274790/
- June 25, 2013, the Bipartisan Policy Center, "A Policy Forum on the Use of Big Data in Health Care"
- "Embracing the complexity of genomic data for personalized medicine." Mike West, Geoffrey S. Ginsburg, Andrew T. Huang, Joseph R. Nevins
- http://en.wikipedia.org/wiki/Omics
- Ibid.
- West M. et al. Genome Res. 2006;16:559-566
- "Embracing the complexity of genomic data for personalized medicine." Mike West, Geoffrey S. Ginsburg, Andrew T. Huang, Joseph R. Nevins
- "The Learning Health System and its Innovation Collaboratives Update Report," Institute of Medicine of the National Academies

---

Todd Saunders, Principal with CBIG Consulting, is responsible for overseeing the delivery of business intelligence, Big Data, and data warehousing solutions and consulting services for CBIG's West Region. Todd has over 23 years of management consulting experience with the last 15 years focusing on business intelligence and data warehousing. Todd began his consulting career with McKinsey & Co. before moving on to Coopers & Lybrand. In previous positions, Todd served as the National Vice President of BI for Braun Consulting and VP of Consulting Services for Quaero Inc. Todd holds a B.S. degree in Physics and Engineering Science from Manchester College, as well as an MBA in Finance and an MSEE in Quantum Electronics from the University of Illinois. Todd is a Certified Business Intelligence Professional and has served as a member of the faculty of The Data Warehousing Institute.