Understanding the Pros and Cons of Big Data Analytics

By Kent Bottles, MD, Edmon Begoli and Brian Worley, PhD

In this article...

Look at how mining big data can lead to medical breakthroughs if the data are analyzed thoroughly.

Big data is a term that is being applied to almost every human endeavor. What is big data? How will the concept of big data affect health care in the coming years? What can physician leaders and hospital system executives do to use big data to decrease per capita cost and increase the quality of the health care they deliver to their patients?

The digitalization of information has created more data and the development of cloud computing, and faster and faster computers have made this increased data more accessible and useful. The insurance company John Hancock's 600 megabytes of data was thought to be the largest amount of data available to any one organization in the 1950s.

By the 1970s, this honor went to Federal Express's 80 gigabytes; during the 1990s, WalMart was believed to have the most data with 180 terabytes. In the early 2000s, Google had accumulated 25 petabytes of data; today most analysts believe Facebook has the most data, an estimated 100 petabytes of data.[1]

The International Data Corp. reported that the amount of digital data exceeded 1 zetabyte in 2010; in 2011 this number was almost 2 zetabytes. Understanding the magnitude of the increase in data is difficult, but this market research firm states there are "nearly as many bits of information in the digital universe as stars in our physical universe."[2] Google's Eric Schmidt claims that every two days we create as much information as we did from the dawn of civilization up until the year 2003.[3]

When you have that much data you can do things differently. In the book *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Viktor Mayer-Schoenberger and Kenneth Cukier define big data as being able to extract new insights and create new forms of value by analyzing large data sets to find actionable correlations. They expand on this insight by writing: "In a big data world...we won't have to be fixated on causality; instead we can discover patterns and correlations in the data that offer us novel and invaluable insights. Big data is about what, not why."[4]

Big data in medicine

Physicians are trained to generate hypotheses that can be tested by the double-blind clinical trial that uses randomization to ensure that the only difference between the control group and the treated group is the therapy or procedure under investigation. Evidence-based medicine focuses on treatments that have survived this rigorous and expensive way of doing things. One drawback to this traditional approach is that experts estimate that only about 25 percent of what doctors do is truly evidence-based.

Big data's focus on correlations, not causality, is difficult for physicians biased toward the biomedical model, where the focus is finding the cause of the disease in order to effectively treat it. "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. In contrast, with a hypothesis, you're never going to be truly surprised at your result," Stanford cardiologist Euan Ashley said.[5]

Atul Butte, MD, PhD, believes that there are actionable correlations that could help patients just waiting to be discovered by data mining of existing health care data sets. "I don't think enough people study the measurements that have already been made. Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world."[6]

The big data approach has already begun to disrupt health care in ways that are only now becoming appreciated. The convergence of genomics, wireless sensors, imaging, information systems, social networks, cloud computing power, and ubiquity of smartphones described by Eric Topol in *The Creative Destruction of Medicine* has given us a glimpse of a new kind of personalized medicine made possible by big data studies.[7]

NAME: MR. PATIENT
AGE: 27 YEARS
DIAGNOSIS: BROKEN ARM
REASON OF ADMISSION: ACCIDENT
CARDIO: REGULAR BEATS, NO MURMURS
PULMONARY: CLAIR, NO RALES
ABDOMINAL: SOFT, NO ORGANOMEGALY
NEUROLOGICAL: NOTHING ABNORMAL DETECTED
X-RAY
BROKEN BONE
LAB FINDINGS:
WBC: 7.5G/L HB: 16.7G/DL
MCV 92 FL PLT: 201 G/L CRP: 0.1 MG/L
NA: 136MMOL/L K: 4.0MMOL/L CA: 2.54
GLYCEMIA RANDOM
CPK 300 U/L
CREATININE: 109 UMOL/L
T: 37C
PULSE: 99

When Stanford researcher Michael Snyder sequenced his own whole genome, he found he was at risk for developing type 2 diabetes. By periodically having his blood tested for about 40,000 proteins, Snyder developed his own integrative personal genomics profile consisting of 30 terabytes of data about how his body functions. The diabetes prediction was surprising because Snyder is slender, has no family history of diabetes, and has never had elevated blood glucose readings. When he consulted with Stanford endocrinologist Sun Kim, Snyder was told, “There is no way you have diabetes.”[5]

Google’s Eric Schmidt claims that every two days we create as much information as we did from the dawn of civilization up until the year 2003. [3]

A three-hour fasting blood sugar test revealed an elevated initial blood glucose level of 127 (normal 70 to 99 mg per deciliter), and later hemoglobin A1C tests were elevated at 6.4 percent and 6.7 percent (normal between 4 and 6 percent) establishing a diagnosis of diabetes.

Snyder has changed his diet and lost 15 pounds; his blood glucose levels have returned to normal. However, for life insurance purposes, he has diabetes and his rates have gone up. Snyder states, “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.”[5] Cardiologist Eric Topol has called Snyder’s integrative personal genomics profile study “a landmark for personalized medicine.”[8]

The Snyder diabetes example exemplifies the transition we are experiencing from a traditional medical paradigm of “diagnose and treat” to a digital personalized medicine paradigm of “predict and prevent.”

Mining the data

A University of Ontario study of sepsis in premature babies is another example of how analysis of large data sets can reveal actionable clinical correlations that are surprising to experienced clinicians.

More data means more difficulty separating the noise from the signal.

By studying the 1,200 data points per second from the wireless sensors attached to the babies, researchers were able to diagnose infections 24 hours before fever and elevated white blood cell count made the disease clinically evident. The actionable correlation revealed by the data was that very steady, constant vital signs indicated impending infection, not well-being.[4]

Studies using the Cancer Genome Atlas (TCGA), a database with more than 300 terabytes attempting to sequence the genome from 20 human cancers, has already changed treatment and classification of some human malignancies. There is a growing appreciation that the traditional classification of tumors by organ of origin (lung cancer, breast cancers, prostate cancers) should be replaced by a classification based on what gene is producing abnormal amounts of protein.

Lung adenocarcinomas with driver mutations for the EGFR gene are clinically responding to the oral medication Gefitiinib, and lung adenocarcinomas with driver mutation for Alk + gene are showing clinical response with Crizotinib. It has also been shown that some breast cancers exhibit the same gene mutations that can be found in some lung adenocarcinomas.[9]

Butte, and Russ Altman, MD, PhD, have developed algorithms that can mine publicly available health information in large databases. Mayo Clinic, Harvard's Partners Healthcare, Vanderbilt, the VA, and Kaiser all have been leaders in establishing large repositories of health information. By combining data from 130 different experiments about gene activity levels in diabetic and healthy tissue, Butte identified a new gene associated with type 2 diabetes because it stood out in 78 of the 130 studies. He estimates that the chance of this result occurring randomly would be less than one in 10 million trillion.

In another paper, Butte used an algorithm looking for drugs and diseases that had opposing effects on gene expression levels to show that the old ulcer drug cimetidine would be affective against some lung adenocarcinomas and that the old drug topiramate could be used to treat Crohn's disease.[6]

Altman used algorithms to mine data from the Stanford Translational Research Integrated Database Environment and FDA's database of adverse-event reports to show that patients taking both antidepressants and thiazide blood pressure medicine are at increased risk for long QT syndrome, a serious cardiac arrhythmia.[10]

Secondary use of data that are not usually considered to be health-related can also be useful for public health purposes. Investigators from Google published an article in *Nature* comparing the 50 million most common search terms in the United States with Centers for Disease Control data on the spread of influenza between 2003 and 2008. By using 450 million different mathematical models, they compared a model of 45 search terms against the CDC documented flu cases in 2007 and 2008. The Google tool successfully predicted the spread of flu a full week earlier than the traditional CDC approach.[4]

The tool known as Google Flu has limitations such as overestimating the impact of flu in 2013 and not helping with tracking new diseases like H1N1 and SARS. The CDC's BioMosaic program is building on Google Flu to combine airline records, disease reports and demographic data that fuel a website and an iPad app that were able to predict that five counties in Florida and five counties in New York were the most vulnerable to cholera spread from the Haiti cholera epidemic.[11]

Big data and waste

Identifying actionable big data correlations can be used by hospital administration to identify and eliminate waste in any department: finance, facilities, billing and pharmacy.

New York City's Office of Policy and Strategic Planning provides examples of how big data analysis can solve real-world problems. Restaurants that illegally dump cooking oil into sewers cause more than half of the city's clogged drains. By looking at publicly available data about which restaurants have contracts with grease recycling firms and geospatial data on the sewers, the analysis provided inspectors a list of likely suspect restaurants; inspectors had a 95 percent success rate in tracking down the offending restaurants.

By asking the right questions of the terabyte of data that flows into the office every day, Director Michael Flowers and his data miners doubled "the city's hit rate in finding stores selling bootleg cigarettes, sped the removal of trees destroyed by Hurricane Sandy, and helped steer overburdened housing inspectors—working with more than 20,000 options—directly to lawbreaking buildings where catastrophic fires were likeliest to occur."[12]

Hospital systems, health insurance companies and medical group practices are all awash in increasing amounts of data. Sophisticated methods now exist that health care leaders can use to decrease per capita costs and increase the quality of the care they deliver.

The recent development of open source big data analytic platforms and the increased affordability of cloud computing solve the expensive problem hospital executives have faced in the past of owning their own data warehouse. In a report by McKinsey titled *Big Data: The Next Frontier for Competition* estimates that such an approach could add $338 billion in value to the health care system.[13]

Frost and Sullivan's *U.S. Hospital Health Data Analytics Market* report states that in 2011 only 10 percent of U.S. hospitals were using data analytic tools.[14] Jeffrey Hammerbacher, the data guru who oversaw data analysis at Facebook, recently left the private sector to work at Mt. Sinai Medical Center in New York because he believes health care is the most important sector of the economy that big data can revolutionize.[15]

Caution needed

Thoughtful hospital system leaders need to be cautious and skeptical about the claims being made for the big data analytics approach to transforming health care. The following experts create misleading expectations for this approach, which is just beginning to be applied to health care.

Chris Anderson's provocatively titled article "The End of Theory: the Data Deluge Makes the Scientific Method Obsolete" makes the claim that large enough data sets can uncover actionable correlations without deep understanding of the problem:

"Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves."[16]

In Big Data: A Revolution That Will Transform How Live, Work, and Think, Viktor Mayer-Schonberger and Kenneth Cukier make a similar point when they write: "Big data is about what, not why" and "datafication represents an essential enrichment in human comprehension." [4]

Automate This: How Algorithms Came to Rule Our World by Christopher Steiner makes algorithms and bots the key players in aggregating information to uncover the actionable correlations. Steiner defines algorithms as a list of instructions that leads the user to a particular answer or output based on the information at hand, and bots as multiple linked algorithms all aimed at performing one task that can roam the Internet in search of data.[17]

In their enthusiasm for the potential of big data, these experts are misleading us about this innovative approach. Alexei Efros is a big data expert with a more nuanced and reasonable attitude; he calls big data "a fickle, coy mistress" fraught with challenges and possible pitfalls.[18]

There has been a backlash against big data predictive analytics that has most recently resulted in an article in *Science* that demonstrated that Google Flu trends had overestimated the number of influenza cases for four years after the original Google article came out in 2009.

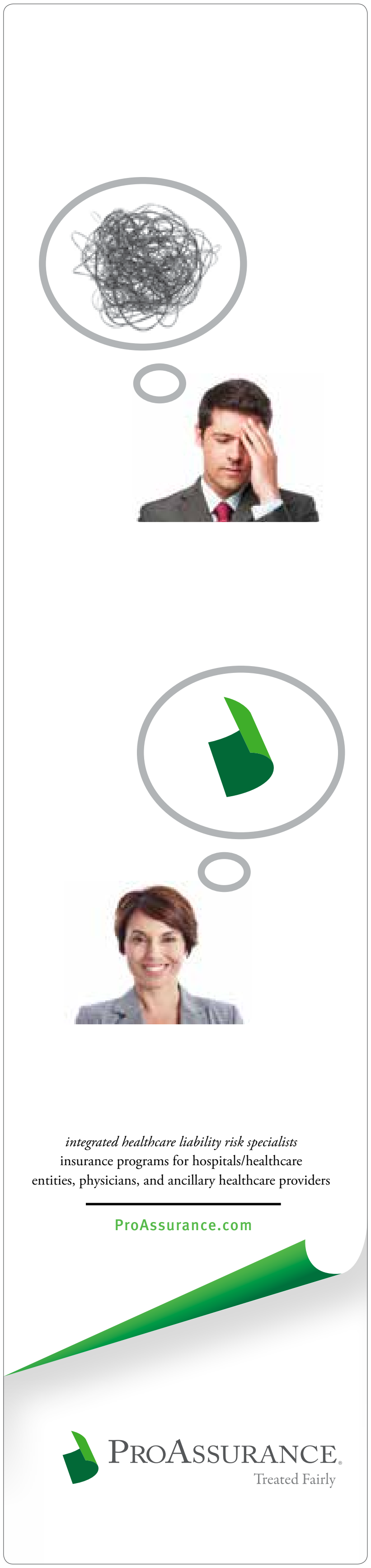
The criticisms of big data predictive analytics include:

1. Inherent biases in how all data are collected and interpreted.
2. Objections to the idea that data mining can replace hypothesis-driven theory by content experts.
3. The observation that the larger the data set the more likely that spurious correlations that are not useful will be identified.[19,20]

David Brooks, Kate Crawford and Lisa Gitelman all forcefully make the important point that there is no such thing as raw data. Gitelman's book *Raw Data Is an Oxymoron* is the most complete analysis of why data are not as objective as we may first think.[21]

As Brooks writes in a column titled "What Data Can't Do:"

"Data obscures values.... One of the points was that data is never raw; it's always structured according to somebody's predispositions and



values. The end result looks disinterested, but, in reality, there are value choices all the way through, from construction to interpretation." [22]

Crawford makes the same important point:

"Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks and are as important to the big data equation as the numbers themselves." [23]

The subjective nature of collecting and interpreting data has convinced us that Chris Anderson's assertion that "with enough data, the numbers speak for themselves" and that the scientific method is obsolete cannot be defended. [16]

Private matters

Big data also presents new challenges to how we try to protect the individual's privacy. The cornerstone of privacy laws has been "notice and consent" where people are told at the time of collection what information is being gathered and the purpose for which it will be used.

The actionable correlations at the center of big data rely on the secondary use of information such as search, wireless sensor data and social media. It is difficult for the individual to give informed consent for secondary uses that are not even imagined when the data are first collected.[4]

Big data presents problems if it's anonymous. While stripping out personal identifiers works in the setting of small data, the reality of big data means re-identification is possible.

When AOL released in 2006 a dataset of 20 million search queries from 657,000 users the information had been carefully scrubbed and was anonymous. By analyzing the searches for items like "tea for good health" and "landscapers in Georgia," *The New York Times* identified user number 4417749 as Thelma Arnold, a 62-year-old widow from Lilburn, GA.

"My goodness, it's my whole personal life. I had no idea somebody was looking over my shoulder," she said.

AOL's chief technology officer and two other employees were fired after the newspaper article was published.[4]

Another pitfall of blindly accepting the big data hype is to ignore the fact that more data means more difficulty separating the noise from the signal. As Nassim N. Taleb points out in *Wired Magazine*, "big data means more information, but it also means more false information."[24]

The idea that big data correlations can completely replace deeper understanding of the cause of problems also requires modification. Evgeny Morozov observes that, "The abandonment of comprehension as a useful public policy goal would make serious political reforms impossible." [25]

That was the experience at the Institute for Clinical Systems Improvement (ICSI) in Minneapolis, MN. In dealing with overutilization of imaging studies, ICSI could not craft a robust and lasting solution until the problem was understood from the point of view of the patient, the primary care physician, the radiologist, the hospital administrator, the employer, and the health plan chief medical officer. Each saw the problem differently, and none saw it in its entirety until a long process of dialogue, give and take, and understanding occurred.[26]

Morozov, in a book review about big data, recognizes that actionable correlations, while useful, may not be sufficient to tackle complicated problems like obesity.

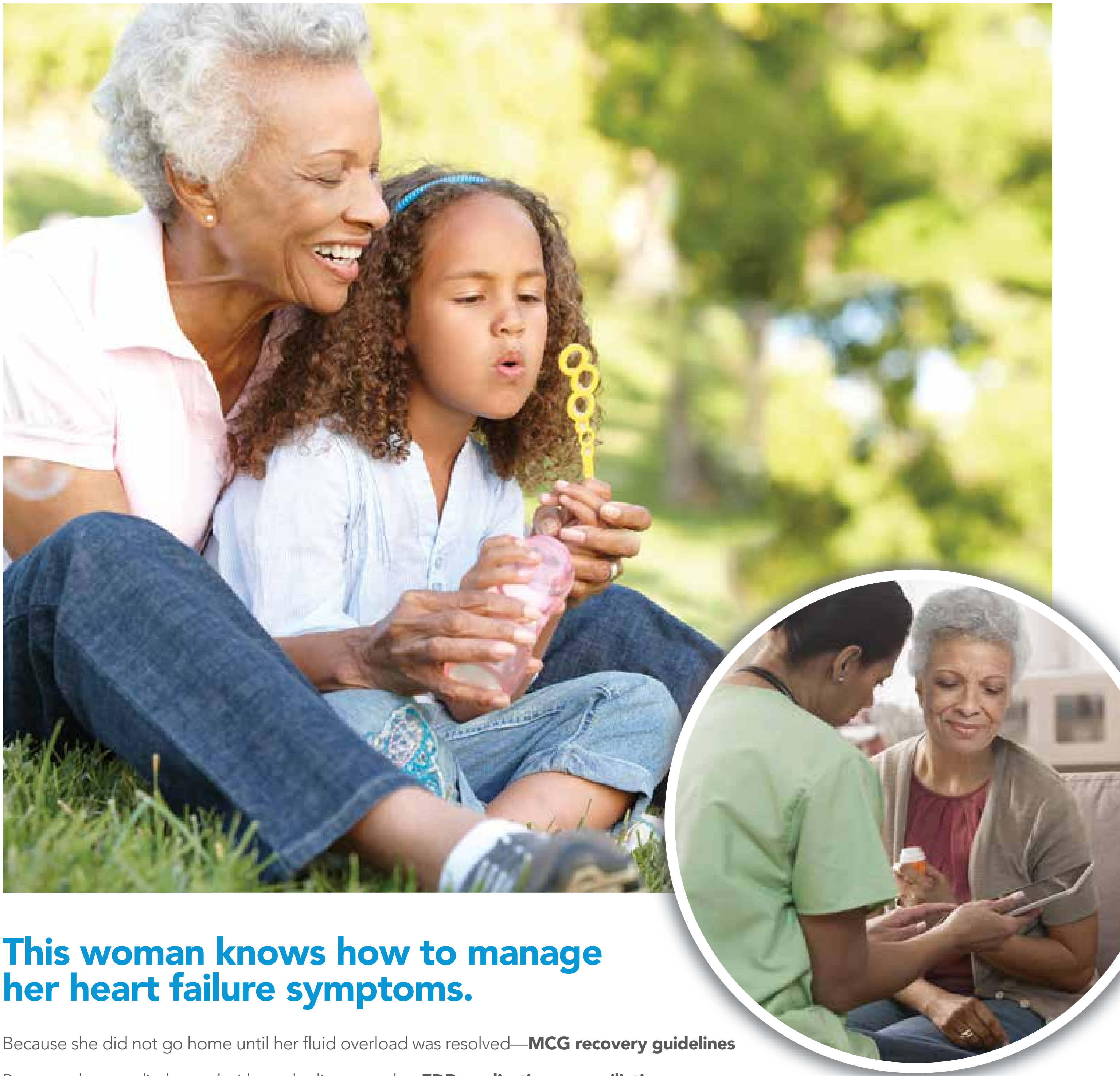
"Take obesity. It's one thing for policy makers to attack the problem knowing that people who walk tend to be more fit. It's quite another to investigate why so few people walk. A policy maker satisfied with correlations might tackle obesity by giving everyone a pedometer or a smartphone with an app to help them track physical activity — never mind that there is nowhere to walk, except for the mall and the highway. A policy maker concerned with causality might invest in pavements and public spaces that would make walking possible. Substituting the why with the what doesn't just give us the same solutions faster — often it gives us different, potentially inferior solutions." [27]

The complexity of dealing with a medical issue like obesity was highlighted in a 2005 *New York Times* article on how class affects cardiac outcomes. The upper class architect Mr. Miele, who did well after his heart attack, had his cardiac exercise rehab program covered by his insurance, and he found a class 10 minutes from his country house.

The less affluent Mr. Wilson lived in a neighborhood without open space to exercise and had to drive into Manhattan during the afternoon rush hour to attend a rehab exercise program. Expensive parking costs made it even more difficult for Mr. Wilson to complete his exercise program.[28]

Derrick Harris strikes the right balance when he acknowledges the potential of big data, but also recognizes the need for human supervision of collection and interpretation of the actionable correlations:

"This is why some people call the process of asking interesting questions of data 'exploratory analytics.' Data analysts can send out a virtual Christopher Columbus to see what's doing inside their data. If they find something potentially valuable, they dig in further. Correlations are just a notice that there might be something worth looking at here.... Machines do the heavy lifting, but humans still play critical roles in training the models by correcting mistakes or adding judgment into an otherwise entirely logical process." [29]



A humorous but serious example of the need for adult human supervision of algorithms is described in a blog post by Michael Eisen, a biologist at UC Berkeley, who wanted to buy Peter Lawrence's *The Making of a Fly: The Genetics of Animal Design* on Amazon's website.

One seller (profnath) had a new copy for $1,730,045.91 and another seller (bordeebook) had a copy for $2,198,177.95. By following the prices for a week, Eisen concluded that one seller was setting his price algorithmically in response to changes in the other seller's prices. The insanity peaked at a price of $23,698,655.93 (plus $3.99 for shipping) until someone noticed the anomaly. The price dropped to $106.23 for one seller and $134.97 for the other seller after human intervention.[29]

Big data is a disruptive technology that will transform health care, and physician leaders would be wise to use data analytics to decrease per capita cost and increase the quality of the care they deliver to their patients.

They also need to recognize the pitfalls and complexity of this new approach. One cannot simply combine multiple databases, crunch the numbers, and magically uncover actionable correlations that can automatically and unthinkingly be implemented.

Human beings with domain expertise and knowledge of the problem being investigated have to oversee the collection of the data, the asking of the right questions, and the interpretation of the results in order to make the best use of this disruptive technology tool.

Kent Bottles, MD, is chief medical officer at PYA Analytics in Knoxville, TN.

kentbottles@gmail.com

Edmon Begoli is chief technology officer of PYA Analytics in Knoxville, TN.

Brian Worley, PhD, is president and CEO of PYA Analytics in Knoxville, TN.

References:

1. online.wsj.com/article/SB10001424127887324178904578340071261396666.html
2. http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm
3. techcrunch.com/2010/08/04/Schmidt-data/
4. Mayer-Schonberger V, Cukier, K. Big Data: *A Revolution That Will Transform How We Live, Work, and Think*. Boston: Eamon Dolan Book/Houghton Mifflin Harcourt, 2013.
5. Conger K. *Big Data*, Stanford Medicine, Summer 2012
6. Goldman B. *King of the Mountain*, Stanford Medicine, Summer 2012
7. Topol, E, MD. *The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care*. New York: Basic Books, 2012.
8. http://www.nytimes.com/2012/06/03/business/geneticists-research-finds-his-own-diabetes.html
9. http://cancergenome.nih.gov
10. Digitale, E. *On the Records*, Stanford Medicine, Summer 2012
11. http://bits.blogs.nytimes.com/2013/06/19/in-new-tools-to-combat-epidemics-the-key-is-context/
12. www.nytimes.com/2013/03/24/nyregion/mayor-bloombergs-geek-squad.html?pagewanted=all&_r=0)
13. http://www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care
14. http://www.frost.com/c/10046/sublib/display-report.do?id=NA03-01-00-00-00
15. http://www.youtube.com/watch?v=OVBZTDREg7c
16. www.wired.com/science/discoveries/magazine/16-07/pb_theory
17. Steiner C. *Automate This: How Algorithms Came to Rule Our World*. New York: Portfolio/Penguin, 2012
18. http://www.newyorker.com/online/blogs/elements/2013/04/steamrolled-by-big-data.html
19. http://www.sciencemag.org/content/343/6176/1203
20. http://www.economist.com/blogs/economist-explains/2014/04/economist-explains-10
21. Gitelman L. (ed). *Raw Data Is An Oxymoron*. Cambridge, MA: The MIT Press, 2013.
22. http://www.nytimes.com/2013/02/19/opinion/brooks-what-data-cant-do.html
23. http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html
24. http://www.wired.com/opinion/2013/02/big-data-means-big-errors-people/
25. http://www.slate.com/articles/technology/future_tense/2013/06/with_big_data_surveillance_the_government_doesn_t_need_to_know_why_anymore.html
26. Scharamer CO. *Theory U: Leading from the Future as It Emerges*. San Francisco: Berrett-Koehler Publishers, 2009.
27. http://online.wsj.com/article/SB10001424127887324178904578342234223307970.html
28. http://www.nytimes.com/2005/05/16/national/class/HEALTH-FINAL.html?pagewanted=all
29. http://gigaom.com/2013/05/28/if-youre-disappointed-with-big-data-youre-not-paying-attention/
30. http://www.michaeleisen.org/blog/?p=358