Can “Big Data” Save a Life?

A Perspective From:
Bob Kelley, PhD
Senior Vice President, Analytics
Truven Health Analytics
March 2013

TRUVEN HEALTH ANALYTICS™

What If?

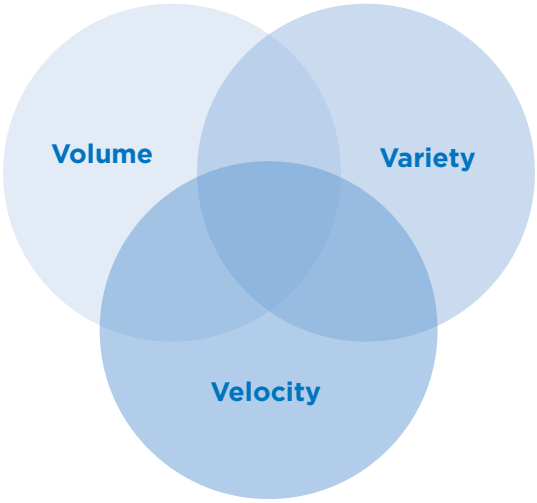
On my next birthday I will reach my father's age when he died of a major heart attack. My dad was, in some ways, like every other World War II veteran. He worked very hard, smoked, and was never sick enough to visit a physician. Upon his death in the hospital, his physician mentioned to me that my dad had likely suffered several less severe attacks and that often men like him simply treated these as heartburn by taking antacids in an attempt to relieve the pain. Until I started to consider the likely application of "Big Data" analytics in healthcare, I treated this as an interesting anecdote to share when the conversation somehow turned to unexpected deaths. It was true that when cleaning out my dad's home we found a variety of antacid containers, conspicuously located within reach of his favorite spots in the living room, on the front porch, and in the bedroom.

Some historians pursue a type of speculation referred to as "counterfactual history." The objective of this method is to speculate on how a single change, such as an alternative decision by a military leader, might have led to a totally different sequence of subsequent events and the eventual outcome. Ever since picking up the book, "What If? 2: Eminent Historians Imagine What Might Have Been" by Robert Cowley at a book sale, I have been fascinated by this exercise in considering the contingencies that affect all of us. You are probably asking yourself, "What do his father's death and interest in counterfactual history have to do with Big Data?"' Let me tell you.

What if my dad had been enrolled in an Accountable Care Organization, perhaps through the Veteran's Administration, which offered its members the service of monitoring all available data relevant to their health and possibly suggesting the need for a healthcare intervention? Of course this slightly bends the rules of practicing counterfactual history, since this option was not available in 1987, and in fact, isn't even available now. But you get the idea. Let's use this scenario to consider the possible value of Big Data to healthcare.

Big Data: The 3 V's

In February of 2001, Doug Laney of META Group Inc. wrote a paper titled, "3D Data Management: Controlling Data Volume, Velocity and Variety." These have come to be known as the 3 V's of Big Data and I find them useful in considering the value of Big Data to specific opportunities for healthcare analytics. I think that visualizing these dimensions as overlapping spaces provides a useful perspective.



For any potential application of Big Data analytics, it is essential to consider the dependence of the method on each of these three dimensions. For example, does the value of an application depend on a real-time analytic that must be inserted in the high-velocity flow of a particular type of data? Or does the application depend on linking a variety of seemingly unrelated types of data from multiple sources? Or does its value depend on having a high volume of data related to a very large population over a long stretch of time? Or does it depend on two or even all three of these characteristics?

Let's consider the implications of each of these three dimensions of Big Data in healthcare analytics, and apply these implications to our counterfactual scenario.

Volume

Let's think about an example described in a recent *Wall Street Journal* article on the data mining techniques applied in this type of pattern discovery. In a 2011 study of the possible side effects of combining two prescription drugs, the authors described using data mining to measure the effect of this combination on blood sugar level. Neither of these drugs alone had a demonstrated effect. In order to confirm this suspected side effect, the authors needed to sift through enough data to find enough cases where a patient had a baseline blood sugar reading and was then prescribed the first drug; then had another blood sugar reading before being prescribed the second drug; and then had a final blood sugar reading. A search of a single medical center's database located only eight patients that met this criterion. After extending the search to two other very large medical centers, a total of 130 cases were identified — enough to complete a reliable analysis.

In our scenario, the doctor's observation that many patients self-diagnose and treat minor heart attacks as heartburn cannot be considered adequate evidence to support a specific intervention intended to prevent a major heart attack for an unsuspecting person. There are many questions that must be answered before this casually observed pattern might be considered for conversion to a reliable and useful aid to patient care decisions:

- How sensitive is an increase in use of antacids in detecting an imminent heart attack?
- How often does an increase actually precede an attack?
- What constitutes a meaningful increase in use of antacids?
- Are there other factors, such as personal characteristics, recent healthcare services and diagnoses, or changes in work environment that have an impact on the answers to these questions?

If the objective is to discover a meaningful and reliable indicator of high probability for a major heart attack in the absence of an intervention (e.g. scheduling the patient for a cardiac catheterization), then these questions must be answered through the application of statistical and data mining techniques. To ensure that these techniques will provide the desired results, they must be applied to a volume of data large enough to include historical data for a sufficient number of people to present the methods with many examples of all possible combinations of antacid use and clinical outcomes. These methods crave volume!

To ensure that these techniques will provide the desired results, they must be applied to a volume of data large enough to include historical data for a sufficient number of people.

How will Big Data analytics applied to new volumes of data enhance current decision-support applications?

Nationally, about 69 percent of Emergency Department (ED) visits could have been served in Urgent Care Centers or even physician offices. The efficiency of the healthcare system could be increased if patients could be encouraged to seek this care from these less resource-intensive facilities. Current decision-support tools identify ED "frequent fliers" using a rules-based measure such as "five ED visits in the last 12 months." Even this simple measure has proven useful to health plans and health systems using Truven Health Analytics™ decision-support tools to target these interventions to those patients most likely to benefit. We can, however, expect that data that links ED utilization patterns with underlying health risk factors (chronic disease), lifestyle choices, and convenience of access to facilities will enable models that reliably identify opportunities to intervene before a patient becomes a "high-flyer."

Big Data provides the opportunity to link and access new data sources that are outside of the healthcare domain, but which might be relevant and provide value to our efforts to better understand the patient and to target interventions to improve health or prevent negative health events.

Variety

To better understand the requirements for and the value of data variety, let's consider two separate categories: domain-specific and disparate type. In the healthcare domain, paid medical claims, lab and other diagnostic results, prescriptions and other orders, vital signs, and patient experience surveys might all be considered domain-specific. Organizations like Truven Health Analytics have significant experience with linking this variety of data sources and preparing the data for useful analyses. Of course collecting and linking this data presents challenges, but these data sources share common attributes that facilitate the use of this data.

Big Data provides the opportunity to link and access new data sources that are outside of the healthcare domain, but which might be relevant and provide value to our efforts to better understand the patient and to target interventions to improve health or prevent negative health events. Innovative organizations are experimenting with these data sources. For example, did you know that the *Journal of Medical Internet Research* exists? In a 2011 article entitled, "Harnessing Context Sensing to Develop a Mobile Intervention for Depression," the authors conclude that, "Mobile phone sensors can be used to develop context-aware systems that automatically detect when patients require assistance. Mobile phones can also provide ecological momentary interventions that deliver tailored assistance during problematic situations." They describe "machine learning techniques that can monitor and learn to recognize a patient's circumstances and state." A personalized, context-sensitive intervention can then be delivered to the person through a mobile phone app.

In our scenario, the need for variety is obvious. We are seeking a signal in retail sales data — a disparate data source — in order to assign risk of a medical event. It is important to acknowledge two serious issues with the access and application of this type of data source.

The first is the individual's right to privacy and the requirement for authorized access to this personal information. The ethical access and use of electronic data is a serious social issue and a personal concern for many in our society. Most people are comfortable with a requirement of a person's consent to access as long as there is a restriction on its use to a specific purpose. There is certainly disagreement on what constitutes real consent, but it is likely that many people will permit access if personal benefit can be demonstrated. It is, therefore, important that this value be documented with real examples of results demonstrating the direct connection between this information and interventions to improve health (e.g. saving a life). I'm sure that my dad would have consented to access of this information, if presented with evidence of relationships between self-medication behavior and life-threatening events.

The second issue is a methodological one. With access to high volumes of data containing a large number of data elements, it is likely that some relationships between a characteristic and an outcome will appear to be significant

when they are not. A recent article in the *Journal of the American Medical Association* (JAMA) highlights this issue while confirming the value of Big Data in observational studies to detect the relationship between characteristics and rare medical events. The authors suggest that, "Few dispute the importance of observational studies for capturing rare adverse events." They describe the difference in results between a randomized controlled trial (RCT) analysis of 14,000 patients with only 284 hip or femur fractures, and only 12 atypical fracture events, across just over 3.5 years of follow-up; and a later observational study that examined 205,466 women for an average of four years with more than 10,000 hip or femur fractures and 716 atypical fractures.

This second analysis, with more than six times the number of the rare event, was able to demonstrate an increased risk of atypical fractures associated with bisphosphonate use that could not be demonstrated by the smaller RCT study. The concern, however, is that such large observational studies cannot adequately control for confounders identified in relationships between variables that may actually result from relationships among any number of other factors in the data set — and not actually be predictive. For example, in our scenario, a relationship between antacids and heart attack may result from a common relationship to a third variable that hasn't been controlled for in the samples. Interpreting these results requires subject matter expertise, offering a plausible medical mechanism for the relationship in our scenario of self-diagnosed heartburn.

Combining dozens of data sources, each at these high levels of velocity, and searching for patterns that suggest some type of intervention in real-time requires new efficient computing and analytic capabilities.

How will Big Data analytics applied to new varieties of linked data enhance current decision-support applications?

Truven Health Analytics health plan and health system clients use existing data sources to identify people who have missed important screening exams, such as a routine mammogram. This data is also used to proactively remind people to schedule the exam in advance. Big Data methods applied to new personal data sources will enable users to better personalize these messages to more effectively encourage them to take care of their health. Machine learning techniques applied to a feedback loop will continuously modify these messages for maximum impact.

Velocity

More than 23 billion credit card transactions are processed in the U.S every year. That's 63 million each day, 2.6 million each hour, and 44,000 every minute. This is high velocity. Intermediaries process more than 1.2 billion fee-for-service claims every year for more than 1 million providers, 2,300 every minute. This is high velocity. Combining dozens of data sources, each at these high levels of velocity, and searching for patterns that suggest some type of intervention in real-time requires new efficient computing and analytic capabilities. These statistics make it clear that incoming Big Data has attained high velocity and it's clear that this trend will continue. However, we are more concerned in this discussion with the velocity of the analyses required to effectively use the information in healthcare decisions, and how to coordinate that with the velocity of the incoming data sources.

In some cases the need to access, link, and analyze these data sources is in real-time, requiring that all sources be updated with the most recent information and that the information provided be available for an immediate decision. An example might be an application to identify risk associated with taking a new prescription drug given other drugs, diagnosed medical conditions, allergies, and diet. Other applications will monitor the receipt of new information on a daily or weekly basis, seeking out patterns with a known health risk and alerting care managers to an increase in risk for a specific person. It is important that the frequency of analysis match the pace of the required intervention.

> **In our scenario,** adequate velocity might be a daily or weekly review of new data to alert the care manager to the importance of scheduling a phone call and possible office visit. It is clear that, in this case, the risk can only be recognized over a period of time sufficient to suggest a measurable increase in the purchase of antacids. The challenge for the methods developed to measure this risk is to enable a timely intervention. If the method accurately measures the increased risk, but cannot reliably enable a timely intervention, it is not useful. Can the increased use of antacids be detected with enough reliability to prevent the event? Of course this depends not only on the risk measure, but on the speed with which the healthcare system can intervene.

How will Big Data analytics applied at high velocity enhance current decision-support applications?

Many health systems apply Truven Health risk models to the retrospective evaluation of outcome measures such as rate of readmission to the hospital within 30 days of discharge. The results of these analyses are used to identify opportunities to change the process of care in an attempt to reduce this rate. However, if we want to reduce the likelihood that a specific patient will be readmitted, information on the patient's risk calculated on data collected and analyzed monthly is not useful. The lag makes any intervention based on this information impossible. Not only do we need to calculate the risk with more velocity, but to obtain the required accuracy in this measurement, we need data from multiple sources to be integrated with the same velocity. With this information velocity, a caregiver can respond with a timely intervention and prevent a readmission before the patient's condition even has a chance to deteriorate.

Veracity: The Fourth Dimension

Truth, accuracy, precision. There is nothing new in Big Data when it comes to the need for data quality. Effective analytics requires that the incoming data be valid and reliable enough to support useful conclusions. What is different is that we expect more from Big Data and the new analytic methods being applied to discover useful information. What's also different is the expectation that accessing new data sources will create opportunities for more value. Both of these characteristics of Big Data analytics create new demands on the methods employed to ensure data veracity, as it specifically relates to the method and its objectives.

We expect more from Big Data and the new analytic methods being applied to discover useful information.

Linking data across sources requires effective algorithms and models for matching individuals and organizations that may be identified differently across sources. After all, if we are trying to measure an individual's risk, it is essential that all of the important data is actually for that individual. Methods for discovering new relationships and for assigning individual risk must recognize the lack of certainty represented by a probability assigned by the matching algorithm. These methods combine the potential error in the risk assignment with the potential error in the matching of the individual across data sources.

Traditional data analysis techniques have assumed that the data is sufficiently structured to ensure consistency in the values of individual records within a single data source and across all data sources. Many of the new and emerging Big Data sources contain little structured data, and efforts to standardize even this structured data across so many data sources is too onerous. Therefore, new methods for extracting useful data from unstructured sources and from non-standardized structured data are required. Subject matter experts are essential to this effort. The danger is that experts in mathematical and statistical methods will attempt to extract information from these sources without the benefit of contextual knowledge.

> **In our scenario,** I have suggested that linking simple grocery and pharmacy retail data to medical claims will provide new insight into consumption behaviors predictive of health events or outcomes. The first challenge is to match individuals across sources. With consent, it should be possible to effectively match using common identifiers like social security number and address. The person can be offered routine opportunities to verify these matches, ensuring a high level of accuracy. Although the retail transactions data is likely to be fairly structured, the formats will vary across the multiple sources. There are no standards for file and record formats. It is likely that most sources will use standard product codes, minimizing the need for data standardization. For example, the Kroger's version of CALCIUM CARBONATE 500mg is coded with an NDC Code of 30142-071-27. Of course the list of codes must be updated as new products become available.

I am setting up a LinkedIn group — Big Data and Healthcare — that you are invited to join as a volunteer willing to search for and share your data for the cause.

I Challenge You to Save a Life!

If you've read this far, you must be somewhat interested in the idea of saving lives with Big Data. I still have a few years before retirement, and could even be persuaded to extend that, if the cause is worthy and the prospects of success are high. I need your help to create a Hollywood-ending for this story (perhaps the role of the son, that's me, could be played by Daniel Day-Lewis). That ending would be the discovery of some reliable signals in Big Data that would trigger an intervention that would prevent a fatal heart attack, and thus save a life. To do this, we need data. I have plenty of medical claims data that provides information on fatal heart attacks and a patient's history of healthcare services prior to the event. What we need are sources of disparate data that can be searched for this elusive signal hidden within the noise. The data might be grocery store or pharmacy sales records, or text messages, or Google searches, or book sales.

I am setting up a LinkedIn group — Big Data and Healthcare at http://linkd.in/15o3NsL — that you are invited to join as a volunteer willing to search for and share your data for the cause. I am quite serious about this challenge and have access to the resources required to collect and manage the data, and to support the type of analyses that can find any signal. Let's make it happen.

References

1 Burns, Begal, Duffecy, Gergle, Karr, Giangrande, Mohr; "Harnessing Context Sensing to Develop a Mobile Intervention for Depression"; *Journal of Medical Internet Research* Vol 13, No 3 (2011)

2 Prasad, Jena; "Prespecified Falsification End Points: Can They Validate True Observational Associations?"; *JAMA*. 2013;309(3):241-242

FOR MORE INFORMATION
Send us an email at **info@truvenhealth.com**
or **visit truvenhealth.com**

TRUVEN HEALTH ANALYTICS™

ABOUT TRUVEN HEALTH ANALYTICS

Truven Health Analytics delivers unbiased information, analytic tools, benchmarks, and services to the healthcare industry. Hospitals, government agencies, employers, health plans, clinicians, pharmaceutical, and medical device companies have relied on us for more than 30 years. We combine our deep clinical, financial, and healthcare management expertise with innovative technology platforms and information assets to make healthcare better by collaborating with our customers to uncover and realize opportunities for improving quality, efficiency, and outcomes. With more than 2,000 employees globally, we have major offices in Ann Arbor, Mich.; Chicago; and Denver. Advantage Suite, Micromedex, ActionOI, MarketScan, and 100 Top Hospitals are registered trademarks or trademarks of Truven Health Analytics.

truvenhealth.com | 1.734.913.3000