Big Data *and* Business Analytics

Edited by
JAY LIEBOWITZ

Foreword by
Joe LaCugna, PhD, Starbucks Coffee Company

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
AN AUERBACH BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20130220

International Standard Book Number-13: 978-1-4665-6578-4 (Hardback)

Library of Congress Cataloging-in-Publication Data

Big data and business analytics / editor, Jay Liebowitz.
pages cm
Includes bibliographical references and index.
ISBN 978-1-4665-6578-4 (hardcover : alk. paper)
1. Business intelligence. 2. Business planning. 3. Decision making--Statistical methods. 4. Data mining. 5. Management--Statistical methods. I. Liebowitz, Jay, 1957-

HD38.7.B54 2013
658.4'72--dc23 2013004216

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

7

Saving Lives with Big Data: Unlocking the Hidden Potential in Electronic Health Records

Juergen Klenk, Yugal Sharma, and Jeni Fan

CONTENTS

What if you could be alerted, perhaps through your smartphone, that you may be about to have a heart attack, stroke, or some other medical event—well before its onset? And that this warning would be based not on commonly recognized symptoms but on a sophisticated data analysis of your vital signs and other health information. In such a scenario, your medical data would be continuously monitored and scanned by powerful computers searching for complex patterns—the patterns of thousands of heart attack or stroke victims, for example, whose pre-event data looked just like yours do now. Alerted to the danger in real time, you could seek emergency medical attention.

While this capability is not yet at hand, advanced research by teams of physicians and data scientists is yielding promising results. In a significant collaboration, Booz Allen Hamilton and a large hospital system in the Midwest set out to find whether a data analysis of past patients' medical

records could help hospitals deal with dangerous, hard-to-treat infections. Their research discovered previously unknown patterns in the historical data that could predict when such infections might suddenly become particularly life threatening.

This is big data—but with a twist. While most analytics rely on the latest available information—to look for emerging business trends, for example—this kind of analysis instead looks backward with big data to try to predict the future. The U.S. government is now at the cutting edge of this approach, developing highly sophisticated techniques to find patterns in past activity that might anticipate threats such as terrorism and cyberattacks on our nation's infrastructure. Other sectors may well find benefit in this approach. Government financial regulators trying to prevent another meltdown, for example, might look at the historical data patterns of banks that failed and see whether similar patterns are emerging in banks today.

In medicine, such an approach could be applied to a host of diseases and conditions—with the potential to save many lives. Just a few years ago, that would not have been possible. But with the rapidly growing transition from paper to electronic health records, vast amounts of medical data are now becoming accessible to researchers. At the St. Louis–based Mercy health system, which collaborated with Booz Allen on the study, Dr. Thomas Hale says that until about five years ago, all of Mercy's patient records were on paper. And collecting data for research was difficult. "To get the data, I had to hire a nurse—and we were lucky if we could collect data on a hundred patients." With the move to electronic health records, he says, "We're now collecting data on three million patients."

Such a wealth of current and historical patient information is one of the key requirements in using data analysis to predict future medical events. If health data analysts are to find critical hidden patterns—if they are to pinpoint clear signals through all the noise—they need as rich a data source as possible.

This search for patterns in data from electronic health records represents a new but valuable tool for physicians. Dr. Hale, the executive medical director of Mercy's Center for Innovative Care, says that traditionally, "Someone comes into our office and gives us symptoms, and we know what the disease is. What we're saying now is, what else is the data showing us that we need to explore? This is entirely different from what we're used to doing as physicians."

Dr. Hale compares the process to that in *Moneyball*, the popular book and movie in which the Oakland A's baseball club achieved success by using computer analysis to find undervalued players. "You take the data and find data points you would not have traditionally suspected," says Dr. Hale.

SURVIVING SEPSIS

The project had its origins in an annual employee ideas contest at Booz Allen, the strategy and technology consulting firm. Among the winning entries of the 2010 contest was the notion that electronic health records might be leveraged to improve the quality of healthcare and patient outcomes. Booz Allen agreed to fund the idea, and the company reached out to Mercy. Booz Allen had previously worked with the 31-hospital system and knew it had large numbers of electronic health records that might be suitable for the research project. Mercy was interested.

The next step was to settle on what disease or condition to research. Booz Allen wanted to pick an area that would have a major impact on Mercy and would provide insights that could be used right away. Clinicians at Mercy suggested studying severe sepsis and septic shock, which are conditions that kill hundreds of thousands of patients at hospitals nationwide each year. Severe sepsis occurs when a localized infection spreads throughout the entire body, causing vital organs, such as the lungs or kidneys, to shut down. Often such infections are hospital acquired, originating when the body's primary barriers are compromised. And because the microorganisms causing the infection may be resistant to common treatments—often due to the widespread use of antibiotics—severe sepsis is notoriously hard to treat. According to the Surviving Sepsis Campaign, a global collaboration by healthcare organizations and professionals, 30 to 35 percent of severe sepsis patients do not survive. Even more deadly is septic shock, which occurs when the organ that fails is the heart. At that stage the patient is typically receiving active treatment in the intensive care unit (ICU)—yet even so, the death rate is about 50 percent.

Dr. Hale likens the progression of sepsis to pouring water into a glass, with severe sepsis occurring when the glass is almost full, and septic shock occurring when the water overflows. The key, he says, is catching sepsis

early. "Once you're septic, you start showing signs and symptoms, and the problem is they're not always picked up in time," he says. "The reason you have such a high morbidity is that you may not catch it in the early stages, when it is tissue inflammation and not organ failure."

Mercy initially wanted to use the data analysis to find out how well its hospitals were complying with treatment "bundles," or protocols, developed by the Surviving Sepsis Campaign. The protocols call for taking certain lab tests and administering antibiotics and fluids—all in a particular order and within a specific time frame. Sepsis treatment is generally not standardized in hospitals across the country. Physicians might order one test but not another, or they might prescribe the antibiotics but not the fluids, or they might take the individual steps out of order or outside the time frame.

Mercy also wanted to know the correlation between compliance with the protocols and patient mortality. That question was critical, because although the protocols had been compiled as best practices by healthcare experts, they had never been systematically tested on a large scale. Such a task would have been extremely difficult with paper health records, because of the need to track the relationship between a number of individual steps that may or may not have been applied to each individual patient.

A second major goal of the research with Mercy was early detection. The analysts wanted to see whether an analysis of data from previous patients, whose conditions had worsened into severe sepsis, might reveal previously unknown patterns in vital signs and other readings. If so, those patterns might be used to identify current patients who were at high risk for severe sepsis.

While both parts of the project called for extracting data from electronic health records, the search for patterns represented new ground for the Booz Allen–Mercy study team. As far as could be determined, this would be one of the first times a data analysis had been performed on electronic health records to try to predict the onset or worsening of a condition or disease. Similar research has since been conducted or is now underway at other facilities, but at the time, the Booz Allen and Mercy analysts and clinicians were on their own.

Although electronic health records offer valuable opportunities for data analysis—they are a far cry from paper records—they offer their own substantial challenges. Most electronic patient records are intended to be read by people, not computers, and do not naturally lend themselves to data analysis. In addition, the sheer volume of information in electronic health records is daunting—each one used by Mercy, for example, has about

8,000 fields in which information can be entered. These fields catalog every last detail of a patient's hospital stay, from symptoms and vital signs to tests, treatments, medications, and a host of other factors that are duly noted along the way. To complicate matters, vendors of health-records software typically establish their own sets of information fields and design particular ways that medical professionals can view the data—such as bar graphs that show the number of patients with a certain diagnosis. This lack of standardization among vendors can make it difficult to compare records among hospitals that are using different systems.

NEW APPROACHES TO GATHERING DATA

To solve such problems, the Booz Allen–Mercy team gathered information from the electronic records in an innovative way. Their approach was drawn from work conducted by Booz Allen in collaboration with the U.S. government. Intelligence analysts searching for terrorists and other threats need the ability to paint a comprehensive picture that considers all kinds of data at once. Booz Allen and the government addressed this problem by developing what is called a *data lake*—a new kind of information repository that is beginning to change the shape of data analysis.

Data lakes represent a completely different mindset from current advanced analytic techniques like data mining. Users no longer need to move from database to database, pulling out the specific information they need. With a data lake, the information from any number of databases is essentially dumped into a common pool, making it easier to ask bigger, more complex questions.

Just as important are the new ways that all of this pooled information can be used. Analysts now typically search for answers by creating limited datasets and then asking specific questions based on hypotheses of what the data might show. A keyword search of a database is a simple example, though the questions can become extremely detailed. If users want to ask different kinds of questions, they often have to reengineer both the databases and the analytics involved—a process that can be prohibitively long and expensive. This tends to limit the complexity of the questions that are asked. Not so with the data lake, which frees users to easily tap all of the data in a variety of constantly changing ways.

Perhaps the most transformative aspect of an analytics architecture that incorporates a data lake is that users do not need to have the possible answers in mind when they ask questions. Instead, they can "let the data talk to them." The ability to make complex inquiries, easily switching in and out any number of variables, allows users to look for patterns and then follow them wherever they may lead. This is particularly important in predictive analytics, when people may not know exactly what they are looking for.

The Booz Allen–Mercy team adopted several of these techniques, though their task was simplified because all of the information used in the study came from a single source, Mercy's electronic records. As similar studies become more complex—using electronic health data from many different sources in all types of formats—comprehensive data lakes will be essential. It would be impossible otherwise to analyze so much varied information—and find the critical patterns within it.

A QUESTION OF TIME

Although the data for the sepsis study came from only one place, there was a great deal of it. The study team collected anonymous data from the electronic health records of 27,000 Mercy sepsis patients from four hospitals over a two-year period. Most had a mild form of the condition, but about 6,000 had advanced to the more life-threatening stages of severe sepsis and septic shock. Of the data fields available in the electronic records, the team chose the most relevant 4,000 for the study—giving them more than a hundred million separate pieces of patient information to work with.

But it was not enough to simply collect the information. Before the analysis could begin, the team needed to establish an *ontology*—or set of organizing principles—for the data. This was needed so that the team could ask questions of the data and get answers in a way that would make sense for the study. Essentially, an ontology gives the raw data its needed context for analysis. This was particularly important here because electronic health records have no inherent organization or context. Each record is just a collection of disparate and often loosely related information about an individual patient.

The study team ultimately chose as the primary organizing principle one that cut across all of the data—time. Each bit of patient information—each

test, each vital sign, each treatment—would be put in chronological order. Such an *event-centric ontology* was a natural choice for the study's goals. Determining whether Mercy's hospitals followed the treatment protocols for sepsis—which called for taking certain steps in the right sequence and time frame—dictated organizing the data by time. The same was true if the team was to determine whether a certain action (precisely following the treatment protocols) led to a certain outcome (a lower mortality rate). And, of course, the team needed to see the data in chronological order to determine whether progression of the condition could be predicted.

Electronic health records themselves are not organized by time. A list of tests given to a patient, for example, will not necessarily be shown in chronological order. However, organizing the data for analysis in this manner was possible for the team because of a key feature of electronic health records—every item entered into a patient's file is electronically stamped with the date and time. Or at least should be, in theory—a small percentage of the data did not have a stamp. A larger challenge lay in dealing with time stamps that were inaccurate. It was not uncommon to see events occurring in an illogical order—time stamps might show blood being drawn, for example, after the patient left the hospital.

The team discovered several reasons why time-stamp problems occurred, including that clocks in different computer systems were not synchronized, or that there was too big a gap between the time that tests and medications are administered and when the information was entered into the system. Such gaps in time logic were flagged automatically during the process of preparing the millions of pieces of data for analysis. Team members resolved some of the discrepancies by talking to the doctors and nurses who had treated the patients, though in other cases information had to be left out. While the study would have been stronger had all the information been usable, the team concluded there was enough data available to have confidence in the study's conclusions.

In preparing the data, one other task was necessary—standardizing the medical language so that drug names, units of measurement, test results, and other information were expressed in a consistent manner. For that, the team leveraged an open-source medical vocabulary software known as SNOMED CT.

The entire process of collecting, preparing, and integrating the data—all before it could be analyzed—consumed the lion's share of the time the team spent on the study. This is typical in data analysis, where many of the most difficult challenges lie in the preliminary spadework.

EVALUATING COMPLIANCE

The study team then began its analysis. The first task step was to determine how well Mercy was following the severe sepsis protocol bundle. Using data from the four hospitals, the team looked at how often all the elements of the bundle were adhered to—that is, whether doctors ordered all the lab tests and treatments, and whether they did so in the prescribed order and time frame. The analysis revealed that this compliance occurred with about 17 percent of sepsis patients. That figure was in line with estimates that compliance at hospitals nationwide is generally under 20 percent.

This part of the analysis also examined the impact that compliance had on patient mortality rates. It found a direct correlation—the greater the compliance with the protocols, the fewer patients died of severe sepsis or septic shock. For example, at the hospital with the lowest compliance—just 10 percent—nearly 60 percent of patients died. At the hospital with the highest compliance, where the protocols were precisely followed about half the time, only about 20 percent of the patients died. While the results were perhaps not surprising, they marked the first time the severe sepsis bundle had been tested through data analysis using electronic health records. What ultimately made this possible was the unique ability to analyze large amounts of patient data in chronological order.

The results had an immediate impact on Mercy. Officials quickly began an initiative to make sure the sepsis protocols were implemented at its hospitals. "When we saw the numbers, it was a wake-up call," says Dr. Timothy Smith, vice president of research at Mercy. "We didn't waste any time—people's lives were at stake."

Smith says one reason for low compliance is that the doctors most familiar with sepsis bundles tend to be in intensive care units, where patients with advanced stages of the condition, severe sepsis and septic shock, are typically treated. Doctors on the hospital floor or in the emergency room do not typically manage advanced cases and so are less familiar with the protocols. However, says Dr. Smith, it is critical that sepsis be recognized and treatment initiated in the condition's earliest stages—before it becomes life threatening and the patient is transferred to the ICU.

In a pilot program at its St. Louis hospital, Mercy educated doctors and nurses on sepsis and the sepsis bundles and took steps to make sure the protocols were implemented in a timely manner—for example, expediting the delivery of antibiotics from the hospital pharmacy.

The results of that effort have been remarkable. Because the protocols are being used earlier and more often, many more patients are surviving the dangerous advanced stages of sepsis. During the first nine months of the pilot program, the mortality rate for patients at the hospital with severe sepsis was cut almost in half—from 28 percent to 14.5 percent. The results for patients with septic shock, which causes heart failure, were even more significant. Prior to the initiative, about 47 percent of septic shock patients died, slightly below the national average. That figure dropped to just 18.5 percent. Mercy estimates that in this initial period alone, the pilot program saved nearly 100 lives. Says Dr. Smith, "We anticipate lives saved to be in the thousands once the program is generalized to all our hospitals."

EARLY DETECTION

Those kinds of outcomes were just what Booz Allen was hoping for when it set out to study how applying data analysis to electronic health records might positively impact patient care. But the research team wanted to take it a step further and see whether even more lives could be saved by actually predicting the severe worsening of sepsis—so that patients could be treated before the condition got out of hand.

For this part of the study, the team examined the data of septic patients whose condition had worsened into severe sepsis. The hope was that advanced data analysis might reveal certain patterns in the data that could serve as red flags. It was here that the team members were asking that the data "talk" to them. Since the analysts didn't know in advance what those red flags might be, they needed to see whether patterns might emerge on their own. This required an entirely new level of data analysis, one more sophisticated than the examination of compliance with sepsis protocols.

The study focused on three key vital signs—heart rate, respiratory rate, and temperature. Here again, the organization of the data by time was critical. A single reading of a vital sign may or may not mean anything. But how vital signs change in combination over time can be far more revealing—that is where crucial patterns begin to emerge.

The team analyzed the progression of the three vital signs of about 1,500 patients who started out with uncomplicated sepsis. About 950 of those patients went on to develop severe sepsis. Were there differences in the progression of vital signs between the patients whose conditions

worsened and those who did not worsen? Could those differences reveal previously unknown red flags that might lead to earlier diagnosis?

During the analysis, several important patterns did in fact emerge. And these enabled the study team to create a computer model that could predict when a patient is at high risk of moving into severe sepsis. The model was preliminary, requiring further development and testing. But it demonstrated that advanced analytics applied to electronic health records could provide insight into the progression of many diseases and conditions.

In practice, patients diagnosed with uncomplicated sepsis are typically already receiving the necessary treatment, and knowing they are at risk of developing severe sepsis may not prompt a different course of action. But the value of the study was that it found indicators the patient might be *worsening,* at no matter what stage of sepsis—that the glass of water, in Dr. Hale's analogy, is steadily filling. Such information is critical for early diagnosis and treatment.

THE NEXT PHASE: CONTINUOUS MONITORING

Early warnings—of sepsis or any other condition—can be fully effective only if patients are continuously monitored. While such monitoring does occur in intensive care units, the vital signs of non-ICU patients around the country are typically taken only once every eight hours, or perhaps once every four hours for patients who need closer observation. One of the frustrating challenges of sepsis is that those time frames are often not enough to catch the condition before it rapidly spins out of control.

Until recently, continuous monitoring on all hospital floors was not practical. However, new technologies, including inexpensive, noninvasive monitoring strips, are now becoming widely available. As part of a new "virtual sepsis initiative," Mercy is beginning to use these monitoring strips to capture real-time, continuous biometric data on patients receiving care in non-ICU beds. In this project, patients hospitalized with simple sepsis or considered to be at risk for sepsis are being monitored for signs that they might be progressing to severe sepsis or septic shock. The idea, says Dr. Smith, is to try to detect such a progression as early as possible and speed the implementation of all the sepsis bundle elements. The patients' doctors are looking not only for the previously known symptoms

of sepsis but also for several of the new indicators that were uncovered by the Booz Allen–Mercy study.

For example, says Dr. Smith, the data analysis revealed that an important indicator may be when the heart rate and respiratory rate go up at the same time—something doctors had not been fully aware of. Although the simultaneous rising of the two rates doesn't by itself indicate sepsis, he says, it does show that the patient is experiencing the kind of distress that sepsis can cause. And it can help alert doctors that a patient not known to have sepsis might have the condition, or that a patient already diagnosed might be worsening into a more severe state.

Advanced analytics does not supplant the doctor's traditional approach, but rather aids it by providing new and perhaps critical information. As Dr. Hale puts it, "We still want to look at the patient heuristically and use our experience. But now here's more information about a patient that will help us make our decision."

INTERPRETING DOCTORS' AND NURSES' NOTES

While the Booz Allen–Mercy study was limited in scope, it laid the foundation for several areas that will require further study. One was the thorny challenge of doctors' and nurses' notes. Such notes often contain important information about patients that do not necessarily appear in one of the data fields—for example, a doctor might write that a patient was sweating profusely or had significant pallor. This kind of "unstructured" information could be valuable to data analysts looking for patterns in patient symptoms, and it often may be needed to gain a full picture. What this means is that if electronic health records are to be used to their full potential, researchers will have to find a way to turn those notes into a format that can be analyzed.

The common approach to translating prose into computer speak is known as natural language processing, but notes from doctors and nurses do not easily lend themselves to this technique. Most natural language processing is designed for complete, properly ordered sentences. Doctors' and nurses notes, in contrast, are typically filled with sentence fragments, medical shorthand, and other quirks such as the framing of patient conditions in the negative—as in, "The patient was *not* sweating."

The study team attacked this challenge by bringing together a variety of natural language processing techniques, many of them developed at academic institutions and placed in the public domain. Team members selected the most suitable techniques and then customized them specifically for use on electronic medical records. Because of the study's time constraints, the team was not able to incorporate enough information from the doctors' and nurses' notes to have a significant impact on the study results. However, the progress made by the study team will help point the way for further research.

TOWARD THE FUTURE

The study's success in finding predictive patterns in historical medical data has important implications for the future of healthcare. As electronic health records become commonplace, large amounts of patient data on virtually every condition and disease will be available for analysis. While initial research is likely to continue to focus on identifying infections in hospitals, data analysts and doctors will eventually be able to aim their sights in almost any direction.

An area that holds particular promise is mobile patient monitoring, which frees doctors to keep an eye on patients out of the hospital setting as they go about their daily lives. Although several forms of mobile monitoring have been widespread for years, they currently do not leverage the kind of data analysis of historical electronic health records that was explored by Booz Allen and Mercy. Matching historical patterns with a patient's continuous readings would greatly expand the ability of doctors to catch and even predict worsening conditions before they turn dangerous.

New kinds of mobile monitoring devices to make this possible are emerging, from wristbands to skin patches to pills that send out data from the gut. The opportunity lies not only in providing better care to the individual patients being monitored but also in analyzing all of these new streams of information—to constantly build and refine even better predictive models.

There are limitations, of course, on the ability to anticipate how and when a patient's condition may change. The further in advance one tries to predict, the lower the accuracy will be. It may not be possible to look around 10 corners, but advanced data analytics may help doctors

look around one or two. And having crucial information about that short time frame may be all that is necessary to save a patient's life, whether in a hospital setting or on the street.

The application of advanced data analytics to electronic health records is just beginning, but early studies, such as the one by Booz Allen and Mercy, show great promise. The healthcare community, which has been adopting electronic records, now faces a new challenge—how to take full advantage of them to benefit patients and reduce medical costs. This challenge was reflected in a provision of the federal stimulus legislation that gives medical care providers financial incentives for the "meaningful use" of electronic health records. The Booz Allen–Mercy study demonstrated how data analytics can help achieve that goal.

It also suggested how the meaningful use of data might be considered in other areas of business. As in healthcare, just amassing big data is not enough—it is what you do with it that counts. "From my standpoint," says Dr. Hale, "if you take the clinical aspect out of it, these are all the various things one wants to do in business." The key, he says, is in "finding data points that are easy to monitor, but that you didn't realize actually had an impact on your business. You use that to improve your business practices and make a positive change."