Big Data in Health Care Decisions: Where Are We Now and What Does the Future Hold?

Marcus Wilson, PharmD, President, HealthCore, Wilmington, DE, USA

KEY POINTS . . .

The evidence base that powers health care analytics has to be sufficient and relevant to the individual patient or patient population to which analytical tools are being applied.

The gap in evidence is enormous and must be effectively addressed if we are to improve quality and affordability; to be sure, the potential of analytic tools to help close this gap is considerable.

It is critical to have a comprehensive view of the individual built through integration of various types of data in order for analytic tools to have the chance of achieving their potential — that means, not just big data, but deep data.

For additional information in this issue:
"Big Data" was the topic presented at the **ISPOR 20th Annual International Meeting** in Philadelphia, PA, USA, as part of the Third Plenary Session: "Big Data, Big Systems, and Better Evidence: What Progress?" (see page 24).

This article is based on a presentation during the First Plenary Session, "The Use of "Big Data" - Where Are We And What Does The Future Hold?" at the ISPOR 19th Annual International Meeting, May 31-June 4, 2014, Montreal, QC, Canada

Few topics ignite more excitement in health care now than the imminent prospect of harnessing big data to better manage, and even cure our myriad illnesses. The forecasted applications of big data range from enhanced evidence development and improved decision-making capabilities to more effective tools for evidence-based population health management. The compelling potential of big data applications has engaged the imagination of the best and brightest in biomedical, clinical, data analytics, information technology and health care business circles and attracted substantial investment and resource commitments. Still, as is common with technological change, big data is enveloped in frenetic research, development and, the tricky part, assimilation into a variety of nuanced health care requirements.

So, the key question at this juncture is: Where are we now in big data computing, and what does the future hold?

To glean some answers, it might be helpful to start with a brief historical perspective. More than 20 years ago, many clinicians responded to their increasing frustration over the lack of applicable evidence by finding new and improved means of generating and disseminating medical evidence. The applicable evidence that clinicians needed included insights at the point of care for clinical recommendations to physicians and patients and to help policy makers grappling with reimbursement and related decisions at the institutional level. Today, applicable evidence is commonly referred to as real world evidence or RWE.

Our goal was to extract diagnostic and treatment information from what was essentially transactional data to gain some insights into patients' real world outcomes. Beginning with an effort to integrate pharmacy and medical claims, which at that time were kept in separate databases, we explored numerous ways to build a robust profile of the patient population enrolled in the Blue Cross/Blue Shield plan in Delaware. Several years later, after the Delaware data were integrated and we developed a workable understanding on how to or when not to use administrative claims data, we began work with a much larger Blue Cross/Blue Shield plan in the state of Florida. Our first assignment was to integrate their pharmacy and medical claims. Fortunately for us too, the Florida health plan had a million lives worth of laboratory results data stored in its basement, still in shipping boxes. The laboratory data had yet to be integrated with any other data, but the plan managers had reasonable expectations that once the data were organized and evaluated important insights on patients' demographic and clinical characteristics as well as treatment outcomes would emerge. So we embarked on integrating their pharmacy, medical, and laboratory results data, which provided an excellent source of real world data for our studies and for analytic insights for a number of years.

During those years, fledgling health outcomes research operations like ours at the time struggled to find usable data and fairly structured sources like the Blues of Florida and Delaware were few and far between. Today, our repository of linked longitudinal medical and pharmacy claims data draws from 14 geographically dispersed US commercial health plans and represents approximately 46 million lives. Our researchers currently have access to electronic laboratory data for large numbers of patients, which are regularly incorporated into their study designs. In addition, we enrich the patient data in our claims database via patient surveys (patient reported information) and, increasingly, by abstracting and reviewing both paper-based medical charts and electronic medical records. How far we have come in just 20 years!

To generate the best evidence we have to start with the best data. We must be able to integrate data from disparate sources to get the best view into the individual's experience >

within the health care system and to power the descriptive, predictive and prescriptive analytics and tools that we want to develop as a discipline.

To be sure, we are by no means singular in data environment growth and expansion. Today, there are very large and growing databases of millions of lives. If added together, these health care data repositories would seem to have data on more than 700 million lives, although the country's population is slightly more than 300 million – likely the result of switching between health plans or double enrollment, the most common being simultaneous coverage under Medicare and a commercial health plan (e.g. Medicare Part D, which covers pharmacy benefits). In tandem with the growing data availability has been a commensurate expansion of the health care analytics industry as a whole. Data from market researchers MarketsandMarkets (Fig. 1) shows that spending on health care analytics has grown significantly in the course of the last 10 years. Furthermore, they indicate that we have just entered the most significant growth phase — the period just prior to the mature phase of the health care analytics sector. This growth is driven by a number of factors:

- Availability of greater volumes of electronic data;
- Advances in analytic methods and improved computational power;
- Concerns with the lack of generalizability of clinical trials [1,2];
- Emphasis by payers on the relevance of study populations and how representative they are of the memberships of their health plans as they make and implement policy decisions;
- Providers in the US increasingly assume more financial risk, along with the clinical risks associated with their decisions; and
- The need for more effective clinical decisions at the point of care.

For a better understanding of the various forms of analytics and their interrelatedness and interdependencies, analytics may be broken down into three different categories: descriptive, predictive, and prescriptive [4]. Descriptive indicates what happened in the past and why, predictive suggests what will happen in the future, and prescriptive specifies what to do about it. In essence, descriptive data provides the evidence base and predictive and prescriptive insights help to put that evidence to work. Without the descriptive foundation, predictive and prescriptive analytics may not only be inexact, they will have little or no reliability. Without ties to predictive and prescriptive analytics, descriptive becomes esoteric. Outcomes of care are best when all forms work collectively to support decisions on a given therapeutic area (Fig. 2).

In applying this model to a disease state, we may start by trying to understand the human health experience and how that experience varies depending on genetic and environmental factors. We then marry that knowledge with what we know about any individual or population of interest to better predict what will likely happen to them (the outcome) over time. If the predicted outcome appears unsatisfactory, we can explore various types of interventions, analyze them, and then apply the results of those analytics in an effort to change the outcome.

As we conduct such outcomes research, it is critical to think about the types of decisions we are trying to empower. This is an essential precursor for thinking in an applied manner, to better measure endpoints and to qualify and communicate our findings to the intended audience.

Today, by and large, each of the three forms of analytics is still done manually. The consequences are several, but the most consequential is curbs on the speed

Figure 1. Growth Trends in the Global Health care Analytics Market [3]

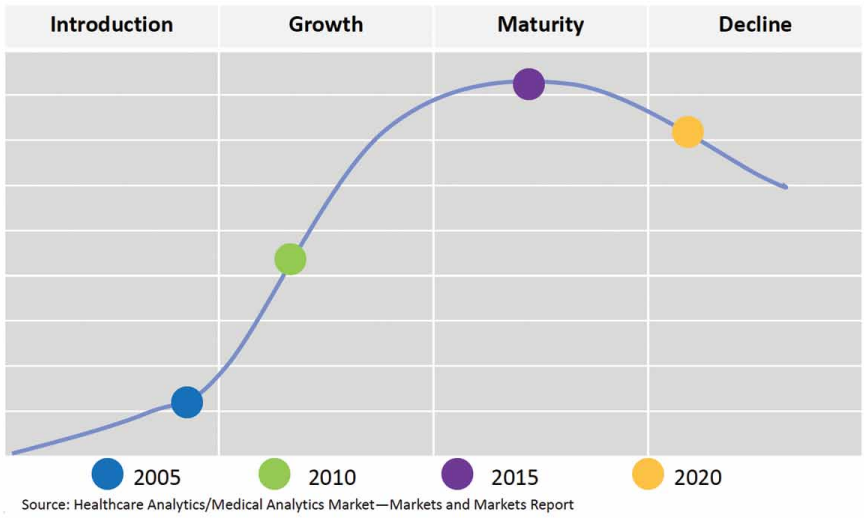


Figure 2. Optimizing Analytic Categories to Improve Patient Outcomes.

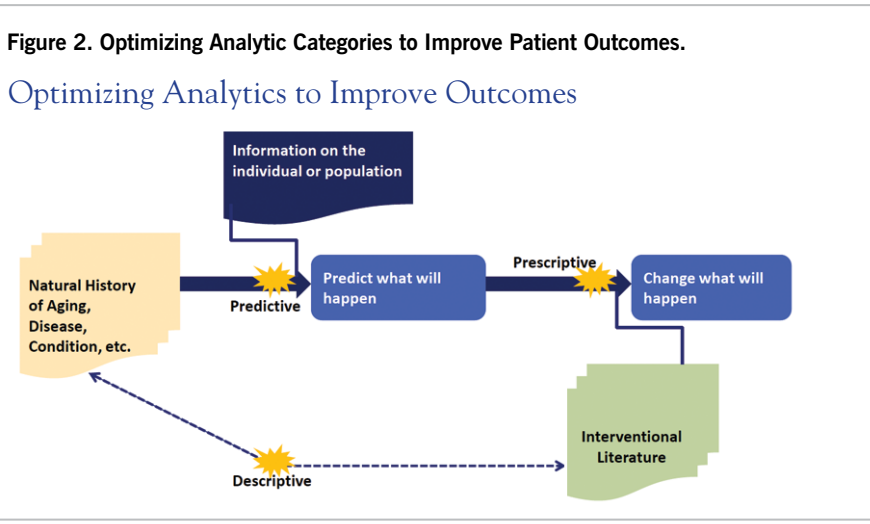
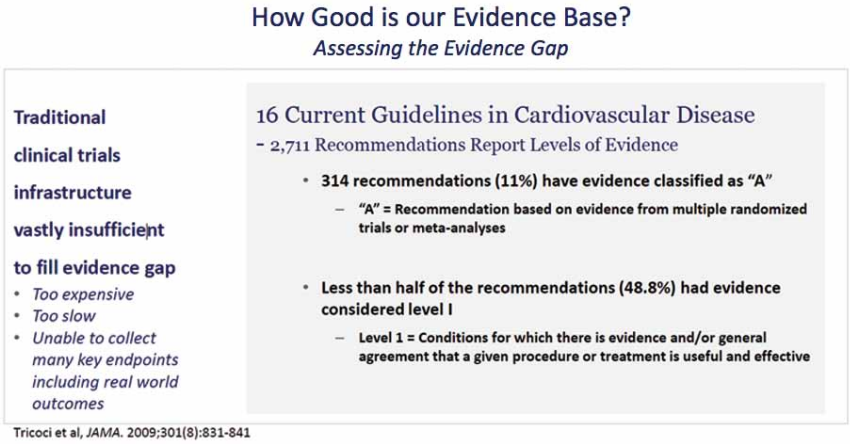


Figure 3. Gaps in the Evidence Base [5].



at which evidence can be generated and findings disseminated to decision makers, especially at the point of care. This inevitably contributes to the long cycle time needed for evidence to be incorporated into practice.

To illustrate this issue, it might be helpful to assess how well we are generating the evidence base for our prescriptive tools. One of the most powerful prescriptive tools in use today is the therapeutic guideline. These guidelines assimilate existing evidence into a familiar format to facilitate clinical decision making on a given condition, patient presentation or therapeutic area. In an effort to better understand the quality of evidence powering these guidelines, Rob Califf and his colleagues at Duke University examined the cardiovascular disease therapeutic guidelines from the American College of Cardiology (ACC) and the American Heart Association (AHA). In this assessment of the 16 current guidelines that report levels of evidence, of the 2,711 recommendations, only 314, or 11%, were based upon evidence classified as Level A (Fig. 3), which means that the vast majority were only supported by a single study or expert opinion [5,6]. These guidelines are a significant step forward, but they clearly illustrate the gap in knowledge that exists in one of the most studied areas of medicine. This is an example of the large existing gap in what we know versus what we need to know to achieve optimal outcomes.

How can we hope to close the quality and affordability gap if we do not have the evidence to guide our decisions? As it exists today, the traditional clinical trial system, while still the mainstay of evidence development is insufficient to generate >

Figure 4. Patients Exposed to Inadequately Assessed Health care Technologies.

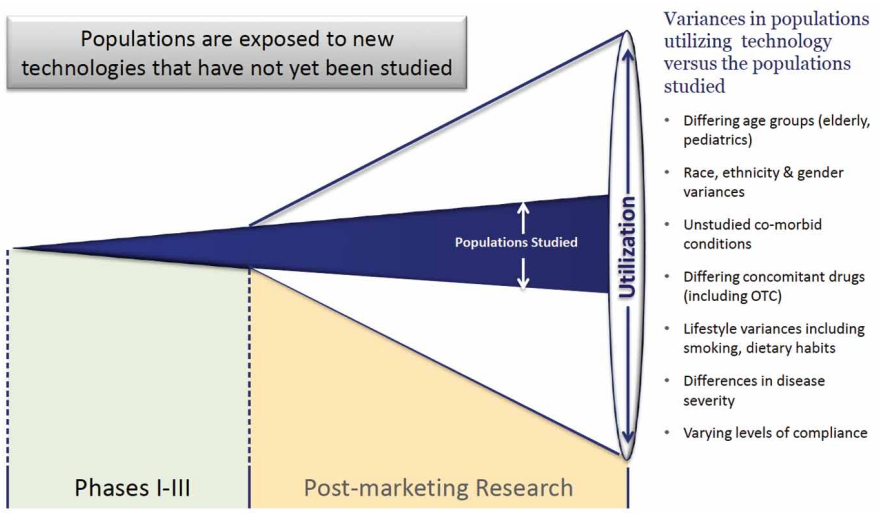
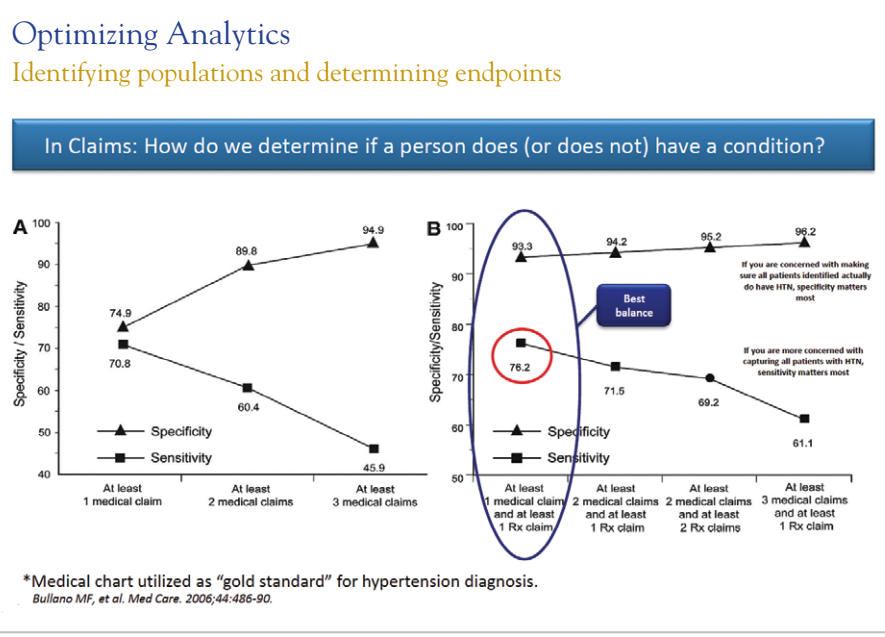


Figure 5. Streamlining Analytics to Identify Populations and Determining Endpoints.



the evidence at the pace needed to close this gap. The system is too expensive and too slow. In addition, all too often the patient populations represented in clinical trials lack sufficient overlap in relevant characteristics to those of patients encountered in clinical practice. This lack of generalizability is the major contributor to the evidence gap (Fig. 4). The end result is population exposure without evidence of value or safety [7-10].

Perhaps the greatest opportunity lies in improving the speed of evidence development on a broad representation of the population in a cost effective manner without compromising the quality of the output. For more than two decades, researchers have experimented with various study designs and data sources including administrative claims, electronic laboratory results, biometric data, health risk assessments and, more recently, electronic health records. The preponderance of research over this timeframe has been done on non-integrated databases typically with data from only one of the foregoing data types represented.

A deep understanding of a given data set (or data type) – its strengths, weaknesses, omissions, inherent assumptions – is critically important. Each data type can have significant limitations requiring adjustments in analytical approaches and techniques. Some examples are provided in Figure 5, which shows the results of a claims data assessment to determine the existence or absence of hypertension diagnoses. To validate the claims analysis, the investigators used patients' medical records as the gold standard, and evaluated a series of combinations of diagnosis and pharmacy codes to determine the sensitivity and specificity of a given combination. A key aspect of this study is the existence of a standard against which the assumptions could be validated. Often data sources do not have a relevant gold standard or appropriate benchmark and, as such, have to be used with caution.

The investigators considered the combination in the oval in Figure 5 to be the best balance of sensitivity and specificity. Of all of the information in the figure, however, the most important number is the one in the circle (76.2%). Under the best-case scenario, using claims data alone, roughly 25% of the patients who had hypertension were not correctly diagnosed. Without knowing that fact, it would be difficult to qualify and study using claims data for an analysis where the existence of a diagnosis of hypertension was important. Unpublished work evaluating the diagnosis of diabetes has shown similar findings.

Claims data are a powerful asset. In fact, due to their ability to serve as an "index" for a patient's experience with the health care system, they are the most effective starting point for an observational study in most cases. However, their limitations must be well understood and the variations that exist between the various claims data sources can have a material impact on a given study [11-15].

We are now at the point where we need to focus on the types of data to be used when attempting to harness big data, and when applying techniques to understand and generate insights from them. The types of data selected and our understanding of that data are critical. The challenges we

now face with marshalling and managing the data in our claims repositories and disease registries could become highly magnified as we seek to incorporate big data into our research environments. After all, big data is akin to a deluge of massive unstructured masses of information too large and unwieldy and not amenable to handling with standard relational database tools — and about 2.5 quintillion bytes are emanating from all segments of society on a daily basis. Even so, there is much riding on our successful ability to tame this raging data tornado—and apply it to the urgent causes of improving patient outcomes and containing runaway health care costs [16-21].

traditional clinical research infrastructure needed to field randomized clinical trials.

To summarize, the analytic market is clearly growing, especially in the areas of predictive and prescriptive analytics. In addition, tools that automate many of these functions are becoming prevalent. As we invest, we cannot forget two critical factors. First, the evidence base that powers these analytics has to be sufficient and relevant to the individual patient or patient population to which the tools are being applied. If we fail to precisely align available data and analytical tools with treatment interventions for the appropriate patient class, it will be quite futile to try to make better decisions

To generate the best evidence we have to start with the best data. We must be able to integrate data from disparate sources to get the best view into the individual's experience within the health care system and to power the descriptive, predictive, and prescriptive analytics and tools that we want to develop as a discipline.

Without a doubt, big data presents challenges as well as abundant opportunities. The greater the depth of the data environment to which you have access, the greater the insight that can be derived from those data. Having access to administrative data is a great starting point, but having the ability to get to both inpatient and outpatient chart data to supplement these data is particularly powerful – and, with the ability to include patient reported outcomes this becomes an even more valuable resource. The greater the overlap in those data sources, the greater the insight. This is a logical statement, almost a truism, but one that needs to be stated as we look into the harnessing and deployment of big data, and especially when incorporating machine learning, cognitive computing and artificial intelligence techniques into our analytic arsenals [22].

Lastly, these integrated data environments are essential components of the future prospective research platform capable of fielding large scale, real world evidence designs such as pragmatic clinical trials. They enable more rapid patient identification for recruitment and greatly reduce the burden on data collection allowing the expansion to more typical treatment settings that do not possess the

to improve quality and affordability. Second, it is critical to have a comprehensive grasp of the types of data and the improvements in the data that we have or will gain access to—that is, not just big data, but deep data. We need to be able to integrate data together to get the best insight in order to both power the analytics that goes into descriptive phases of the analytic work as well as to power the prescriptive and predictive tools that we want to develop as a discipline.

References

[1] Califf RM, DeMets DL. Principles from clinical trials relevant to clinical practice: Part II. Circulation2002;106:1172-5. [2] Califf RM, DeMets DL. Principles from clinical trials relevant to clinical practice: Part I. Circulation 2002;106:1015-21. [3] Healthcare Analytics/Medical Analytics Market by Application (Clinical, Financial, & Operational), Type (Predictive, & Prescriptive), End-user (Payer, Provider, HIE, ACO), Delivery Mode (On-premise, Web, & Cloud) - Trends & Global Forecasts to 2020. Available at: http://www.marketsandmarkets.com/Market-Reports/healthcare-data-analytics-market-905.html. [Accessed August 7, 2014]. [4] Davenport TH. Analytics 3.0. Harvard Bus Rev December 2013:65-71. [5] Tricoci P, Allen JM, Kramer JM, et al. Scientific evidence underlying the ACC/AHA clinical practice guidelines. JAMA 2009;301:831-41. [6] Califf RM, Platt R. Embedding cardiovascular research into practice. JAMA 2013;310:2037-8. [7] Tufts Center for the Study of Drug Development. Lack of Clinically Useful Diagnostics Hinder Growth in Personalized Medicines. Impact Report, 13 (July/August 2011):4. [8] Brass EP. The gap between clinical trials and clinical practice: the use of pragmatic clinical trials to inform regulatory decision making. Clin Pharmacol Ther 2010;87:351-5. [9] Califf RM, Platt R. Embedding cardiovascular research into practice. JAMA 2013;310:2037-8. [10] Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials. 1998; New York: Springer-Verlag. [11] Chastek BJ, Oleen-Burkey M, Lopez-Bresnahan MV. Medical chart validation of an algorithm for identifying multiple sclerosis relapse in healthcare claims. J Med Econ 2010;13:618-25. [12] Lewis NJ, Patwell JT, Briesacher BA. The role of insurance claims databases in drug therapy outcomes research. Pharmacoeconomics 1993;4:323-30. [13] Nathan H, Pawlik TM. Limitations of claims and registry data in surgical oncology research. Ann Surg Oncol 2008;15:415-23. [14] Weiner MG, Lyman JA, Murphy S, Weiner M. Electronic health records: high-quality electronic data for higher-quality clinical research. Inform Prim Care 2007;15:121-7. [15] Zhang J, Yun H, Wright NC, et al. Potential and pitfalls of using large administrative claims data to study the safety of osteoporosis therapies. Curr Rheumatol Rep 2011;13:273-82. [16] Manyika J, Chui M, Brown B, et al. Big data: the next frontier for innovation, competition, and productivity. New York (NY): McKinsey Global Institute; c2011 [cited 2013 Jun 1]. Available at: http://www.mckinsey.com/insights/business_ technology/big_data_the_next_frontier_for_innovation. [Accessed August 16, 2014]. [17] Stonebraker M. What does 'big data' mean? [Internet]. New York (NY): ACM; 2012 [cited 2013 Jun 1]. Available at: http://cacm.acm.org/blogs/blog-cacm/155468-what-does-big-data-mean/fulltext. [Accessed August 16, 2014]. [18] Bellazzi R. Big data and biomedical informatics: a challenging opportunity. Yearb Med Inform 2014;9:8-13. [19] Berger ML, Doban V. Big data, advanced analytics and the future of comparative effectiveness research. J Comp Eff Res 2014;3:167-76. [20] Brennan N, Oelschlaeger A, Cox C, Tavenner M. Leveraging The Big-Data Revolution: CMS Is Expanding Capabilities To Spur Health System Transformation. Health Aff (Millwood) 2014;33:1195-202. [21] Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. Health Aff (Millwood) 2014;33:1178-86. [22] Steve Lohr. IBM's Virginia Rometty on Leadership and Management. Available at: http://bits.blogs.nytimes.com/2014/05/13/ibms-virginia-rometty-on-leadership-and-management/. [Accessed August 17,2014]. ■

Additional information:
To view Dr. Wilson's presentation, please visit the Released Presentations page for the 19th Annual International Meeting at: http://www.ispor.org/EventInformation/2014Montreal?p=212