Unlocking the Power of Unstructured Data

WHITE PAPER

Sponsored by: IBM

Susan Feldman
Judy Hanover
Cynthia Burghard
David Schubmehl

June 2012

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.935.4445 F.508.988.7881 www.idc-hi.com

IDC HEALTH INSIGHTS OPINION

Electronic health records (EHRs) and computerized physician order entry (CPOE) systems capture structured data in the course of clinical care that can be drawn into analytics applications. Structured data aims to create a consistent view of all patients in a healthcare system, providing a predictable set of data that can be tracked and analyzed.

However, while structured data supplies the "what" of a disease or treatment, it rarely can offer the reasons behind decisions.

In many cases, unstructured text remains the best option for providers to capture the depth of detail required, for example, in a clinical summary, or to preserve productivity by incorporating dictation and transcription into the workflow. Unstructured text records contain valuable narratives about a patient's health and about the reasoning behind healthcare decisions.

In the past, unstructured data presented a formidable obstacle to analytics, but new techniques and technologies have helped unravel unstructured data. Contextual search and natural language processing (NLP) applications can help derive meaning from unstructured data and pull valuable medical history elements into the analytics environment. This white paper explores the use of advanced content and predictive analytics to incorporate valuable information from unstructured data into the healthcare organization, providing a more complete view of each patient as a result. Key findings include the following:

- Unstructured data is a valuable portion of the medical record and can be leveraged in analytics programs.
- Textual and predictive analytics tools can allow previously unused content to be made available to analytics in order to improve healthcare for individual patients as well as to uncover patterns of cause and effect and indicators of disease that were previously unknown. The result is healthcare improvement as well as a reduction in hospital readmissions — a significant cost savings to the healthcare system. Providers should use clinical documentation

systems that merge information from structured and unstructured data in order to get a more complete picture of their patients and their treatment patterns.

- Healthcare systems that use content and predictive analytics are discovering that the systems can extract data from unstructured information more consistently than having clinicians fill in forms — a savings in clinicians' time and an overall improvement to data consistency.

IN THIS WHITE PAPER

This white paper is presented by IDC Health Insights and IDC's Search and Discovery Technologies research practice and sponsored by IBM. The objectives were to gain insights into:

- The value of analyzing and mining unstructured data
- The changes and opportunities that mining unstructured data will create for healthcare organizations
- How healthcare organizations are already using this new approach to reveal insights in patient care

METHODOLOGY

IDC Health Insights analysts collaborated with analysts in IDC's Search and Discovery Technologies research practice to explore the use of text mining technologies in healthcare. This white paper was prepared using existing IDC research and analysts' experience as well as customer interviews.

SITUATION OVERVIEW

Why Care About Unstructured Data?

For decades, healthcare organizations have lacked access to data and the tools required to make critical business decisions. The healthcare industry has lagged behind other industries in its use of its existing data for decision making. Progress has been made with administrative and financial data, including billing and claims data, but results have lagged with regard to the information in patient clinical records. Like organizations in other industries, most healthcare organizations maintain multiple legacy systems with the same purpose. In fact, nearly 70% of health plans have multiple claims systems.

This heterogeneity of data sources creates a huge challenge for healthcare analytics, even when only the structured information is being amassed. The growing availability of clinical data, much of it in

the form of unstructured information (text-based results and transcribed dictation), not only adds to the challenge of integrating multiple sources of data but also creates new opportunities for understanding patients, both as individuals and as part of a population.

Only in the past few years has clinical data become more widely available through the deployment of electronic medical records (EMRs), and IDC Health Insights predicts that half of all hospitals will be using an EMR by the end of 2012. While EMR adoptions result in a significant increase in structured data, they also generate significant unstructured data.

Traditional healthcare analysis has focused on structured data, while unstructured information is still in its infancy with regard to its use and incorporation in analytics. However, it is clear that in order to arrive at the depth of understanding, accuracy, and transparency they need from their analytics environments, healthcare organizations will need to integrate unstructured data into this analysis.

Challenges of Unstructured Data

Healthcare organizations are faced with the need to respond to meaningful use and accountable care initiatives that focus on the triple goal of increasing the quality of care, improving patient safety, and reducing costs. New regulations require attention to and progress on improving quality metrics in the course of care delivery, and providers who do not meet goals will see reimbursement rate penalties. These quality and performance management initiatives include areas such as reducing preventable readmissions and achieving meaningful use measures or Health Effectiveness Data and Information Set (HEDIS) measures.

Traditionally, calculating quality metrics or gathering data for research without electronic-form clinical data required laborious, resource-intensive chart review. Providers need to scrub data for errors in data entry and recording, and these efforts are significant.

But data alone cannot reveal attitudes and judgments. Assessing the quality of care requires a full understanding of the clinical documentation. Manual collection and analysis of written records is no longer an option as the volume of information grows. Furthermore, manual analysis of text is often colored by the biases of researchers.

For over a decade, other industries, such as finance and manufacturing, have increasingly relied on content or text analytics to uncover what their customers are saying. Government intelligence uses these technologies to track terrorists or uncover wasteful fraud.

The premise is simple: Language contains patterns, and that's how people learn to speak; computers are good at understanding patterns, so it should be possible to teach a computer to recognize patterns in text and, like a person, to look up meanings in order to "understand" it.

Natural language processing (NLP) processes language the way a person would, examining the sentence structure as well as the words in a document to derive meaning. Content analytics takes that analyzed text and extracts meaningful elements such as names of people, places, drugs, and diseases, as well as their relationships to each other: Which drug causes what side effects? What symptoms are related to which diseases? Merging this knowledge with structured data enables healthcare professionals to discover patterns and relationships across their legacy repositories and the published literature, giving them a more complete picture of their patients and of the treatment patterns or the spread of diseases. It lets them discover quickly which treatments are most effective for which patients — something that no one physician could discover without months or years of research.

Content analytics uses linguistic techniques to determine the meaning of words, using dictionaries, grammatical patterns, and context. At a relatively low level, NLP can extract terms, no matter how they are expressed, to fill in forms automatically. For example, *meaningful use* requirements include "smoking status," and while a query might find the words *smoking* and *smoker* or *smoked*, it does not look further to find *tobacco* or *4 cigarettes per day*. Content analytics and NLP systems may also be trained to differentiate among *former smoker*, *current smoker*, and *heavy* or *light smoker*. While EHRs today have added these contextual fields, NLP can be used to extract this data from historical records or physician notes. NLP systems can also be used to unite multiple databases without having to rewrite all schemas to a single central schema.

There is a more compelling reason, though, for moving to content analytics and NLP: They help healthcare organizations make more fully informed decisions. Information contained in the context of clinical data, such as linkages between side effects and medications, cause and effect relationships and correlations, or demographic information, is rarely contained in traditional databases, but it can be extracted from text.

WHAT DO WE MEAN BY CONTENT ANALYTICS?

Content analytics relies on a series of modules that extract meaning and structure from the multiple layers of text. As a patient record is processed, it is analyzed for the words and phrases that it contains, as well as the patterns — sentence structure, paragraphs, sections, titles, headings — that give it meaning. A series of analyzers tag and extract meaningful elements for analysis. Some of these are:

- Names of people, symptoms, drugs, and diagnoses
- Concepts or ideas so that the same idea can be found no matter how it is expressed (i.e., SNOMED CT codes, *high blood pressure* versus *hypertension*)

- Time (When did something happen? What happened before or after an event?)
- Relationships such as cause and effect, side effects, or other patterns such as the correlation between smoking and lung cancer or exercise and longevity
- Categories so that similar documents, patients, or diseases will be grouped together
- Sentiment or opinion (How did the patient feel about his or her hospital stay: positive, negative, or neutral?)
- Location (Where did something happen? This is important both to speed up emergency medical response and to discover what relationship a location might have to the spread of disease.)

Content analytics uses these building blocks of meaning to create collections of information that can be mined very much as data is mined. While standard search engines treat each document as a separate item, these text mining tools look across electronic collections to forage for patterns, cause and effect, or relationships among the entities in the collection. This powerful set of capabilities will serve as the basis for a new generation of information access and analysis systems that improve population health, discover factors for readmission, predict infection outbreaks, or improve quality measures. While this set of systems focuses on healthcare, other classes of applications using the same kinds of content analytics will help find fraud in insurance claims, deduce where and when terrorists might strike next, and even help predict election outcomes by analyzing social media.

Today, knowledge workers can't keep up with the variety, volume, and velocity of information that bombards them constantly and still do their jobs. Physicians are expected to be omniscient. The promise of these applications is that they will digest and mine the information stack, pulling out what is pertinent, when it is needed to support care givers and health systems in improving quality of care.

CASE STUDY

Seton Healthcare Family

Background

Seton Healthcare Family is a provider of healthcare services in central Texas that operates 38 facilities that serve 1.9 million residents in an 11-county area. The volume of information collected in the EMR across facilities necessitated an automated approach to derive the most understanding possible from its systems. Seton therefore embarked on a new, data-driven approach that would transform patient care. The provider knew that it had a wealth of data that could help it recognize

patterns of disease, uncover new best practices, and, in the end, improve patient health while cutting costs. This new approach, though, was unproven, so Seton started with a pilot project that involved personalized longitudinal management of high-risk patients.

Simultaneously, IBM's Watson technology was receiving a great deal of publicity for successfully beating top *Jeopardy!* winners. Seton's CEO began talking with IBM executives about the potential for using Watson in healthcare and how such technology could be used to enhance Seton's transformation process.

Building a Watson for healthcare, however, takes time, and Seton had a pressing problem: how to detect and predict cases of congestive heart failure (CHF) that would result in readmission to the hospital.

IBM's Content and Predictive Analytics is complementary to IBM Watson and let Seton get started immediately. This solution combines NLP with predictive analytics to harvest and analyze structured and unstructured healthcare data to predict, in this case, readmission factors and the patients at highest risk for readmission. Data showed that 50% of CHF patients are routinely readmitted within 30 days, resulting in stress on patients as well as financial and clinical strains for the hospital. Seton needed to reduce preventable readmissions of CHF patients in order to meet its goals of reducing costs and improving the quality of care. Together, IBM and Seton set out to address this problem.

Seton staffed this initiative through its Analytics and Health Economics department, which consulted closely with the clinician team. Today, this group consists of 20 individuals, including those with programming expertise as well as senior researchers. Two groups exist within Analytics and Health Economics: One group is responsible for routine reporting for Seton, and the other group supports the analytic needs of the organization as they relate to clinical research and business process analytics. A third group is being staffed to provide medical economic research to look at cost effectiveness, cost benefit, longitudinal patient care, and other health plan–like analytics.

IBM created a database that included 36 months of data from Seton's electronic health record system, its cost data, and information from its administrative database. The results included 5,018 encounters that were woven together to create longitudinal patient records. Based on clinical knowledge and a review of the medical literature for the causes of readmissions for congestive heart failure, a list of over 100 predictors was identified and tested. By looking at structured and unstructured data, the two companies identified 18 top predictors of readmissions, and IBM built a model to determine the influence of each predictor in terms of its propensity for readmission. Unstructured data elements such as Medicaid participation, drug and alcohol use, and living arrangement (whether the patient was living alone or not) were identified as leading predictors of readmissions for congestive heart failure.

Results

The results of this pilot project surprised both Seton and IBM:

- 113 possible predictors of CHF were isolated initially, but the list grew as new insights were found. The data that was expected to be most useful turned out not to be. Instead, the system discovered 18 more accurate predictors of CHF readmission. The final 18 consisted of insights not previously known, including the top predictor.

- The structured data, which they expected to be the gold standard for accuracy, instead was found to be less accurate than the unstructured data.

- The unstructured data was more complete and more accurate than the structured data.

Seton defined the success of this initiative as understanding the ability of unstructured data to more accurately identify patients with the highest risk of readmissions for heart failure and determining if accessing unstructured data would be worthwhile. Both goals were met.

Seton was impressed with the quality and completeness of the unstructured data and what it believes to be more accurate detail of the clinical and social behavior aspects of care.

The results were so impressive that Seton has begun to examine its use of structured data versus unstructured data and its approaches to data entry and workflow. The heavy investments made in structured data in the EHR have not yielded anticipated results in many cases, as provider workloads and productivity demands make large amounts of structured data entry unreasonable for clinicians. Seton is exploring the finding that its structured data sometimes lacked completeness and accuracy, whereas its unstructured data was richer and higher quality when extracted with the new techniques.

While many aspects of clinical documentation are best served by structured data entry, there is also an opportunity to apply text analytics to optimize workflows around the capture of unstructured data.

Opportunities

As pilot projects at Seton Healthcare Family and at other health organizations demonstrate, using content and predictive analytics can save lives while cutting costs. As the bar continues to rise for healthcare organizations to improve both operational and clinical outcomes, unstructured data will be critically important.

Reducing preventable readmissions is a clear target for providers, as one in five patients suffers from preventable readmissions and the financial impact is significant; hospitals can expect to be financially penalized for not reducing their rate of preventable readmissions for their Medicare patients.[1]

Seton Healthcare Family's work combining NLP and predictive modeling to identify factors that influence the likelihood of readmissions for patients discharged with CHF represents an application of contextual and predictive analytics that, as the hospital operationalizes the results, will have clinical and operational benefits.

Next Steps

Executives at Seton are in the process of determining next steps for applying NLP with predictive analytics to additional disease states, as they seek to operationalize the information gleaned from their pilot.

FUTURE OUTLOOK

Content and predictive analytics also have clear applications to not only preventable readmissions but also many other high-cost and quality problems providers face, such as nosocomial infections and chronic disease management. Analyzing unstructured information to identify biomarkers and demographic and social factors will deepen care teams' understanding of chronic disease patients and allow accountable care organizations to target patients with disease management programs to head off costly acute episodes.

In short, the incorporation of unstructured information into predictive analytics has the potential to enhance the ability of providers to control preventable readmissions and target high-risk patients, thereby enhancing their competitiveness and profitability in the accountable delivery environment.

While the Seton Healthcare Family case study demonstrates the effectiveness of using content and predictive analytics to solve a specific healthcare problem, preventable readmissions, it also has wider implications for healthcare delivery. By using content analytics to identify biomarkers and risk factors for conditions or scenarios that were previously unknown, healthcare organizations can move from acute to preventive care, improving the health of patients and intervening at the earliest opportunity to prevent conditions or improve outcomes. The results of this new approach will be significant.

[1] http://www.healthleadersmedia.com/content/COM-263665/3-Readmissions-to-Reduce-Now

IDC sees this particular application as an early example of the coming wave of search-based applications that use unstructured information, content analytics, and traditional search capabilities to provide predictive answers to difficult problems. These types of unified information access applications are becoming more visible and popular. According to *Unified Access to Information: Less Seeking, More Finding* (IDC #227780, April 2011), 54% of respondents were looking for applications that combined their structured and unstructured information. Respondents also indicated their top reasons for investing in this type of solution:

- Lower the cost of managing and analyzing information
- Unify access to all information sources
- Provide faster and easier access to legacy data systems
- Provide decision support and management
- Normalize and relate information across repositories

Future systems utilizing the same building blocks will become available to provide "virtual assistants" for identifying fraudulent insurance claims, analyzing customer satisfaction, predicting recurrence of hospital visits, etc. Content analytics is a key enabling technology used to identify and extract information nuggets, patterns, and trends from unstructured information, building up knowledge bases consisting of evidence gleaned from text such as medical articles, clinical diagnoses, and other evidentiary information.

Future search-based applications will also make extensive use of the analysis and visualization capabilities that today's best business intelligence applications offer. Graphical visualizations combined with the unstructured analysis and predictive assessments of systems such as the Seton Healthcare Family application will change the face of many disciplines, including healthcare.

Ultimately, these types of systems will serve to advance well beyond current systems that do not offer this rich base of evidentiary information and will herald a new class of applications that will seek to use all available information, whether it is structured or unstructured.

ABOUT IDC HEALTH INSIGHTS

IDC Health Insights provides research-based advisory and consulting services that enable healthcare and life science executives to:

- Maximize the business value of their technology investments
- Minimize technology risk through accurate planning

- Benchmark themselves against industry peers
- Adopt industry best practices for business/technology alignment
- Make more informed technology decisions and drive technology-enabled business innovation

IDC Health Insights provides full coverage of the health industry value chain and closely follows the payer, provider, and life science segments. Its particular focus is on developing and employing strategies that leverage IT investments to maximize organizational performance. Staffed by senior analysts with significant technology experience in the healthcare industry, IDC Health Insights provides a portfolio of offerings that are relevant to both IT and business needs.