What is Data Mining in Healthcare?

By David Crockett, Ryan Johnson, and Brian Eliason

> Many industries successfully use data mining. It helps the retail industry model customer response. It helps banks predict customer profitability. It serves similar use cases in telecom, manufacturing, the automotive industry, higher education, life sciences, and more.

Like analytics and business intelligence, the term data mining can mean different things to different people. The most basic definition of data mining is the analysis of large data sets to discover patterns and use those patterns to forecast or predict the likelihood of future events.

That said, not all analyses of large quantities of data constitute data mining. We generally categorize analytics as follows:

- Descriptive analytics—Describing what has happened
- Predictive analytics—Predicting what will happen
- Prescriptive analytics—Determining what to do about it

It is to the middle category—predictive analytics—that data mining applies. Data mining involves uncovering patterns from vast data stores and using that information to build predictive models.

Many industries successfully use data mining. It helps the retail industry model customer response. It helps banks predict customer profitability. It serves similar use cases in telecom, manufacturing, the automotive industry, higher education, life sciences, and more.

Data mining holds great potential for the healthcare industry. But due to the complexity of healthcare and a slower rate of technology adoption, our industry lags behind these others in implementing effective data mining strategies.

In fact, data mining in healthcare today remains, for the most part, an academic exercise with only a few pragmatic success stories. Academicians are using data-mining approaches like decision trees, clusters, neural networks, and time series to publish research. Healthcare, however, has always been slow to incorporate the latest research into everyday practice.

“In fact, data mining in healthcare today remains, for the most part, an academic exercise with only a few pragmatic success stories. Academicians are using data-mining approaches like decision trees, clusters, neural networks, and time series to publish research. Healthcare, however, has always been slow to incorporate the latest research into everyday practice.”

The question that leading warehouse practitioners are asking themselves is this: how do we narrow the adoption time from the bench (research) to the bedside (pragmatic quality improvement) and affect outcomes?

The Three Systems Approach

The most effective strategy for taking data mining beyond the realm of academic research is the three systems approach. Implementing all three systems is the key to driving real-world improvement with any analytics initiative in healthcare. Unfortunately, very few healthcare organizations implement all three of these systems.

The three systems are:

❶ The Analytics System

The analytics system includes the technology and the expertise to gather data, make sense of it and standardize measurements. Aggregating clinical, financial, patient satisfaction, and other data into an enterprise data warehouse (EDW) is the foundational piece of this system.

❷ The Content System

The content system involves standardizing knowledge work—systematically applying evidence-based best practices to care delivery. Researchers make significant findings each year about clinical best practices, but, as I mentioned previously, it takes years for these findings to be incorporated into clinical practice. A strong content system enables organizations to put the latest medical evidence into practice quickly.

❸ The Deployment System

The deployment system involves driving change management through new organizational structures. In particular, it involves implementing team structures that will enable consistent, enterprise-wide deployment of best practices. This system is by no means easy to implement. It requires real organizational change to drive adoption of best practices throughout an organization.

If a data mining initiative doesn’t involve all three of these systems, the chances are good that it will remain a purely academic exercise and never leave the laboratory of published papers. Implementing all three enables a healthcare organization to pragmatically apply data mining to everyday clinical practice.

Pragmatic Application of Data Mining in Healthcare—Today

When these principles are in place, we have seen clients make some very energizing progress. Once they implement the analytics foundation to mine the data and they have the content and organizational systems in place to make data mining insights actionable, they are now ready to use predictive analytics in new and innovative ways.

One client is a health system trying to succeed in risk-based contracts while still performing well under the fee-for-service reimbursement model. The transition to value-based purchasing is a slow one. Until the flip is switched all the way, health systems have to design processes that enable them to straddle both models. This client is using data mining to lower its census for patients under risk contracts, while at the same time keeping its patient volume steady for patients not included in these contracts. We are mining the data to predict what the volumes will be for each category of patient. Then, the health system develops processes to make sure these patients receive the appropriate care at the right place and at the right time. This would include care management outreach for high-risk patients.

With another client, we are mining data to predict 30-day readmissions based on census. We apply a risk model (based on comorbidity, severity score, physician scoring, and other factors) to patients in the census, run the data through regression analysis, and assign a risk score to each patient. The health system uses this score to inform which care-path patients take after discharge so that they receive the appropriate follow-up care.

> We are mining data to predict 30-day readmissions based on census. We apply a risk model (based on comorbidity, severity score, physician scoring, and other factors) to patients in the census, run the data through regression analysis, and assign a risk score to each patient. The health system uses this score to inform which care-path patients take after discharge so that they receive the appropriate follow-up care.

Although these predictive models require a committed cross-functional team (physicians, technologists, etc.) and need to be tested over time, these clients are happy with the progress and preliminary results. They are moving beyond the theory of data mining into real, pragmatic application of this strategy.

Using Analytics to Track Fee-for-service and Value-based Payer Contracts

Let's go into more depth about how one of these clients is using data mining and predictive analytics to address a major trend in healthcare today: effecting a smooth transition from fee-for-service (FFS) to a value-based reimbursement model.

We all know that the transition to value-based purchasing is happening. It represents the future of healthcare. But this shift isn't a switch that can be flipped overnight. Instead, health systems must juggle both care delivery models simultaneously and will likely have to do so for many years to come.

> We are working with a team at a large, nationally recognized integrated delivery network (IDN) that is using data mining to help navigate this transition—working to succeed in risk-based contracts while still performing well under the fee-for-service reimbursement model.

We are working with a team at a large, nationally recognized integrated delivery network (IDN) that is using data mining to help navigate this transition—working to succeed in risk-based contracts while still performing well under the fee-for-service reimbursement model. This means that they need to lower their census for patients under risk contracts, while at the same time keeping patient volume steady for patients not included in these contracts.

Monitoring and Predicting Fee-for-service Volumes

A significant percentage of this IDN's revenue comes from out-of-state referrals to its top-rated facilities. The team wants to ensure that these FFS contracts remain in place and supply a steady stream of business. To monitor this process, they have implemented an enterprise data warehouse (EDW) and advanced analytics applications. The EDW aggregates multiple data sets—payer, financial, and cost data—and then displays dashboards of information such as case mix index (CMI), referral patterns for each payer, volumes per payer, and the margins associated with those payers.

This system enables the team to mine data viewing trends in volume and margin from each payer. At this point in the implementation, the team is able to see within a quarter—rather than after a year or two—that referrals from a certain source are slowing down. They can then react quickly through outreach, advertising, and other methods.

> The IDN is an accountable care organization (ACO) with shared-risk contracts that cover tens of thousands of patients. Just as they are bringing referrals into the hospital, they are optimizing care to keep their at-risk population out of the hospital.

As you can see, this innovative system we're developing is still one that is reactive—though it identifies trends quickly enough that the health system can react before the margin takes much of a hit. But we are currently refining the system to become one that is truly predictive: one that uses sophisticated algorithms to forecast decreases in volume or margin by each referral source.

Participating in Shared-risk Contracts

Of course, at the same time as they work to optimize referral volumes, the health system's team must also manage patients in value-based contracts. The IDN is an accountable care organization (ACO) with shared-risk contracts that cover tens of thousands of patients. Just as they are bringing referrals into the hospital, they are optimizing care to keep their at-risk population out of the hospital. They are, therefore, also using the EDW to help ensure that patients in this population are being treated in the most appropriate, lowest-cost setting. Analytics enables the team to monitor whether care is being delivered in the appropriate setting, identify at-risk patients within the population, and ensure that those patients are assigned a care manager.

Health systems nationwide are feeling the pressure of figuring out how to straddle the FFS and value-based worlds until the flip is switched. Having the data and tools on hand to predict their volumes and margins—while managing value-based contracts using the same analytics platform—is giving a significant advantage.

Pragmatic Application of Data Mining to Population Health Management

Another client is using the flexibility of its EDW to concurrently pursue multiple population health management initiatives on a single analytics platform. We are working together on two initiatives that employ the EDW, advanced analytics applications, and data mining to drive better management of the populations in the health system's clinics.

> The first initiative mines historical EDW data to enable primary care providers (PCPs) to meet population health regulatory measures. This clinic's PCPs must demonstrate to regulatory bodies that they are giving the appropriate screenings and treatment to certain populations of patients.

Data Mining to Improve Primary Care Reporting

The first initiative mines historical EDW data to enable primary care providers (PCPs) to meet population health regulatory measures. This clinic's PCPs must demonstrate to regulatory bodies that they are giving the appropriate screenings and treatment to certain populations of patients. Their focus to date has been on A1c screenings, mammograms for women over 40, and flu shots. The EDW and analytics applications have enabled the PCPs to track their compliance rate and to take measures to ensure patients receive needed screenings.

The Health Catalyst® Advanced Application for Primary Care shows trending of compliance rates and specific measurements over time. So, the clinic can view how a patient's A1c or LDL results are trending. They also see patients who may still be in a healthy range but over the last 18 months are trending closer and closer to an unhealthy result, then proactively address the issue.

A fun story from this clinic involves a Nurse Practitioner who joined the practice 20 years ago with a dream of changing the standard of care for diabetes. She tried to create concise reports but ran into one roadblock after another and finally resorted to spreadsheets mapped to EMR fields as a reporting mechanism, realizing it's a less-than-ideal stopgap. Finally, after 20 years, her dream came true with the Health Catalyst solution to deliver monthly reports to individual physicians showing their diabetic patients and respective compliance to the standard of care.

Having this data readily on hand has also enabled the clinic to streamline its patient care process—enabling front-desk staff and nurses to handle screening processes early in a patient visit (which gives the physician more time to focus on acute concerns during the visit). This approach allows physicians to see more patients and devote more time to those patients' immediate concerns. And it allows each member of staff to operate at the top of his or her license and training.

> To better risk stratify the patient populations, we applied a sophisticated predictive algorithm to the data. Using the data, we identified the clinical and demographic parameters most likely to predict a care event for that specific population.

Data Mining to Predict Patient Population Risk

The second initiative involves applying predictive algorithms to EDW data to predict risk within certain populations. This process of stratifying patients into high-, medium- or low-risk groups is key to the success of any population health management initiative. Interestingly, some patients carry so much risk that it would be cheaper to pre-emptively send a physician out to make a house call rather than waiting for that patient to come in for a crisis appointment or emergency room visit. The clinic needed to be able to identify these high-risk patients ahead of time and focus the appropriate resources on their care.

To better risk stratify the patient populations, we applied a sophisticated predictive algorithm to the data. Using the data, we identified the clinical and demographic parameters most likely to predict a care event for that specific population. We then ran a regression on the clinic's historical data to determine the weight that should be given to each parameter in the predictive model.

By applying such a tailored algorithm to the data, the clinic has been able to pinpoint which patients need the most attention well ahead of the crisis. Importantly, the clinic has integrated this insight into its workflow with a simple ranking of priority patients. This allowed for development of improved processes for managing the care of at-risk patients. For example, each week the physicians and care coordinators discuss the risk level of each patient with an appointment scheduled for that week. They can then create a care management plan in advance to share with the patient during the visit.

The clinic also looks at Patient Activation Measure® (PAM) scores and uses that data to determine patient engagement and activation. This leads to shared decision-making between the PCP and the patient, as the physician is able to determine ahead of time those patients who are at higher risk for non-compliance or might be unable to fully participate in their care.

One important aspect of creating a predictive algorithm is getting feedback from clinical experts. Using an algorithm to make an impact in today's care—which is our goal—requires buy-in from the clinicians delivering care on the frontlines.

Data Mining to Prevent Hospital Readmissions

Reducing 30- and 90-day readmissions rates is another important issue health systems are tackling today. We have used data mining to create algorithms that identity those patients at risk for readmission.

When your health system has an adequate historical data set—i.e., you have adequate data about patients with certain conditions who were readmitted within 30 or 90 days—you can mine that data to create an accurate predictive algorithm. The following is a high-level description of steps to learn from a historical cohort and create an algorithm:

1. Define a time period (the parameters of the historical data).
2. Identify all of the patients flagged for readmission in that time period.
3. Find everything those patients have in common (lab values, demographic characteristics, etc.).
4. Determine which of these variables has the most impact on readmissions. You can do this mathematically using a variety of statistical models.

An Introduction to Training Predictive Algorithms

The process of building and refining an algorithm based on historical data is called training the algorithm. We typically use about two-thirds of the historical cohort to train the algorithm. The other third is used as a test set to assess the accuracy of the algorithm and ensure that it isn't generating false positives or negatives.

One important aspect of creating a predictive algorithm is getting feedback from clinical experts. Using an algorithm to make an impact in today's care—which is our goal—requires buy-in from the clinicians delivering care on the frontlines. For them to own the algorithm, trust the data, and incorporate new processes into their workflow, incorporating their feedback is critical.

> In an ideal situation, health systems would have all of the historical data they needed, would train the algorithm, and would quickly start using predictive analytics to reduce readmissions.

The More Specific the Algorithm, the Better

Rather than train an algorithm specific to cases like heart transplant or heart failure, many organizations rely on all-cause or general readmissions data to predict readmissions. However, most of these generic algorithms are only about 75 percent accurate. It's a start, but it isn't enough.

The extra effort to train an algorithm based on a specific population—say, a cardiac population—will jump the algorithm above 90 percent accuracy. If you can define a very specific problem or population—and identify the characteristics unique to that population—the algorithm will always be better. You can use a generic algorithm as a starting place, but to be truly successful you will need to add factors specific to defined populations.

Readmissions in the Real World: A Health System's Improvement Initiative

In an ideal situation, health systems would have all of the historical data they needed, would train the algorithm, and would quickly start using predictive analytics to reduce readmissions. In the real world, things can be a little bit messier. Health systems don't always have the historical data they need at the outset. Sometimes the health system has to improve documentation first and build up the necessary data before launching predictive analytics.

That was the case with one of our health system clients. Rather than starting to implement predictive algorithms immediately, they used an EDW and advanced analytics applications to begin a readmissions initiative with only general readmissions baselines to guide them.

This client decided to begin by focusing on a specific cohort: heart failure (HF) patients. We worked with them to create a data mart for their HF population so they could track readmissions rates and assess how the quality interventions they implemented affected those rates.

> Data mining can also help this health system streamline its efforts by evaluating the relative efficacy of each best practice. For example, if a case manager only has time to apply some of the interventions to a patient, which intervention or combination of interventions will have the most impact?

The health system's team gathered best practices from the medical literature and decided to use interventions that included:

- Medication review. Clinicians are required to review medications with HF patients at discharge.
- Follow-up phone calls. A nurse calls to check that the patient is following the health regimen appropriately (within seven days for high-risk cases and 14 days for other cases).
- Follow-up appointment scheduled at discharge.

As they implemented these best practices, the data flowed into the EDW, and the team was able to see:

- How compliant clinicians were in using the best practices.
- How these best-practice interventions affected 30- and 90-day readmissions.

They also began to add other best practices to their set of interventions.

At that point, the client hadn't yet set up an algorithm to predict risk. Rather, they relied on physicians to flag patients as high risk. Because the health system needs to refine its processes and ensure that the right amount of resources are being devoted to high-risk and rising-risk patients, the team is now turning its attention to predictive algorithms as a method for streamlining its processes and making more effective interventions.

Data mining can also help this health system streamline its efforts by evaluating the relative efficacy of each best practice. For example, if a case manager only has time to apply some of the interventions to a patient, which intervention or combination of interventions will have the most impact?

This brief case study is illustrative of what applying data mining in the real world is all about. If the health system had waited until its stars were perfectly aligned before getting started on its initiative, it might still be waiting today. Perhaps that's why data mining so often doesn't make it out of the academic lab and into everyday clinical practice. But this is the type of effort that is required—the determination to iterate step by step in a process of continuous quality improvement.

> This brief case study is illustrative of what applying data mining in the real world is all about. If the health system had waited until its stars were perfectly aligned before getting started on its initiative, it might still be waiting today.

Data Mining in Healthcare Holds Great Potential

As stated earlier, today's healthcare data mining takes place primarily in an academic setting. Getting it out into health systems and making real improvements requires three systems: analytics, content, and deployment, along with a culture of improvement. We hope that showing these real-world examples inspires your team to think about what is possible when data mining is done right.

Resources

- 4 Essential Lessons for Adopting Predictive Analytics in Healthcare http://www.healthcatalyst.com/predictive-analytics-healthcare-lessons
- Prescriptive Analytics Beats Simple Prediction for Improving Healthcare http://www.healthcatalyst.com/prescriptive-analytics-improving-health-care
- Quality Improvement in Healthcare: Start With Your Healthcare Data http://www.healthcatalyst.com/quality-improvement-in-healthcare-start-with-healthcare-data
- The Best Approach to Healthcare Analytics http://www.healthcatalyst.com/best-healthcare-analytics-approach
- Clinical Data Warehouse: Why You Really Need One http://www.healthcatalyst.com/clinical-data-warehouse-why-you-need-one
- How to Prepare for Value-based Purchasing in 4 Steps http://www.healthcatalyst.com/prepare-for-value-based-purchasing
- Advanced Applications http://www.healthcatalyst.com/advanced-applications/
- Four Levels of Health Activation http://www.insigniahealth.com/solutions/patient-activation-measure
- A Best Way to Manage a CMS Hospital Readmission Reduction Program http://www.healthcatalyst.com/healthcare-data-warehouse-hospital-readmissions-reduction
- How to Sustain Healthcare Quality Improvement in 3 Critical Steps http://www.healthcatalyst.com/sustain-healthcare-quality-improvement

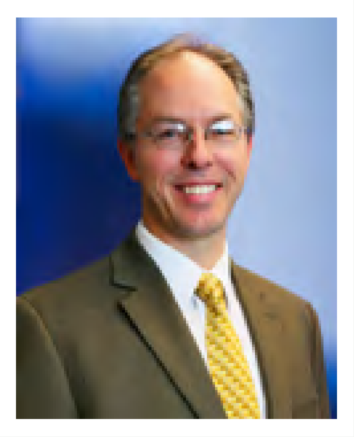
ABOUT HEALTH CATALYST

Based in Salt Lake City, Health Catalyst delivers a proven, Late-Binding™ Data Warehouse platform and analytic applications that actually work in today's transforming healthcare environment. Health Catalyst data warehouse platforms aggregate and harness more than 3 trillion data points utilized in population health and ACO projects in support of over 22 million unique patients. Health Catalyst platform clients operate 96 hospitals and 1,095 clinics that account for over $77 billion in care delivered annually. Health Catalyst maintains a current KLAS customer satisfaction score of 90/100, received the highest vendor rating in Chilmark's 2013 Clinical Analytics Market Trends Report, and was selected as a 2013 Gartner Cool Vendor. Health Catalyst was also recognized in 2013 as one of the best places to work by both Modern Healthcare magazine and Utah Business magazine.

Health Catalyst's platform and applications are being utilized at leading health systems including Allina Health, Indiana University Health, Memorial Hospital at Gulfport, MultiCare Health System, North Memorial Health Care, Providence Health & Services, Stanford Hospital & Clinics, and Texas Children's Hospital. Health Catalyst investors include CHV Capital (an Indiana University Health Company), HB Ventures, Kaiser Permanente Ventures, Norwest Venture Partners, Partners HealthCare, Sequoia Capital, and Sorenson Capital.

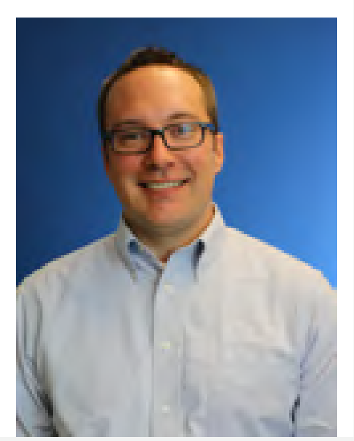
Visit www.healthcatalyst.com, and follow us on Twitter, LinkedIn, Google+ and Facebook.

About the Authors



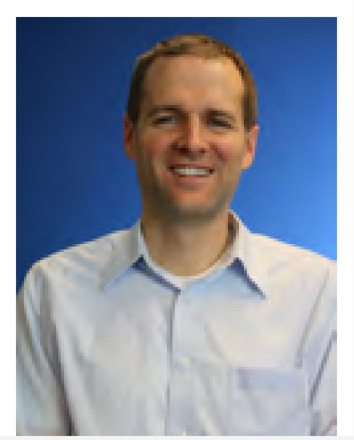
David Crockett

David K. Crockett, Ph.D. is the Senior Director of Research and Predictive Analytics. He brings nearly 20 years of translational research experience in pathology, laboratory and clinical diagnostics. His recent work includes patents in computer prediction models for phenotype effect of uncertain gene variants. Dr. Crockett has published more than 50 peer-reviewed journal articles in areas such as bioinformatics, biomarker discovery, immunology, molecular oncology, genomics and proteomics. He holds a BA in molecular biology from Brigham Young University, and a Ph.D. in biomedical informatics from the University of Utah, recognized as one of the top training programs for informatics in the world. Dr. Crockett builds on Health Catalyst's ability to predict patient health outcomes and enable the next level of prescriptive analytics – the science of determining the most effective interventions to maintain health.



Ryan Johnson

Ryan Johnson joined Health Catalyst in June 2012 as a Senior Data Architect. Prior to coming to HC, he worked 6 years as a software developer for a government contractor, Fast Enterprises, in Utah and Colorado. Ryan has a degree in Mathematics (number theory) from BYU.



Brian Eliason

Brian Eliason brings more than 10 years of Healthcare IT experience to Health Catalyst, specializing in data warehousing and data architecture. His work has been presented at HDWA and AMIA. Prior to coming to Catalyst, Mr. Eliason was the technical lead at The Children's Hospital at Denver with experience using I2B2. Previously, he was a senior data architect for Intermountain Healthcare, working closely with the disease management and care management groups. Additionally, he helped Intermountain bridge clinical programs with the payer-arm, Select Health. Mr. Eliason holds an MS in business information systems from Utah State University and a BS from Utah Valley University.