HealthIT Analytics article – Semantic Computing, Predictive Analytics Need Reliable Metadata

July 18, 2016 – As the healthcare industry reaches the saturation point of electronic health record adoption, and slowly moves past the pain of the implementation process, it may seem like the right time to stop thinking so much about hammering home basic data governance principles for staff members and start looking at the next phase of health IT implementation: the big data analytics environment.

After all, most providers are now sitting on an enormous nest egg of patient data, which may be just clean, complete, and standardized enough to start experimenting with population health management, operational analytics, or even a bit of predictive risk stratification.

Many healthcare organizations are experimenting with these advanced analytics projects in an effort to prepare themselves for the financial storm that is approaching with the advent of value-based care.

The immense pressure to cut costs, meet quality benchmarks, shoulder financial risk, and improve patient outcomes is

causing no small degree of anxiety for providers, who are racing to batten down the hatches before the typhoon overtakes them.

While it may be tempting to jump into quick-win analytics that use “good enough” datasets to solve a specific pressing use case, providers may be at risk of repeating the same mistakes they made with slapdash EHR implementations: creating data siloes, orphaned reports, and poor quality datasets that cannot be used in a reliable, repeatable way for meaningful quality improvements.

Big data analytics, and especially advanced semantic computing or machine learning projects, require more than just a hastily scrubbed-up dataset to help providers navigate the myriad requirements involved in delivering optimal care.

Not only do they rely on clean and accurate data, drawn from the EHR, claims, or other sources, but they require meaningful metadata, as well.

“Metadata” is a set of information that describes another piece of data, such as the timestamp on a prescription authorization, a record of the author of a clinical note, or the size and word count of a document. These pieces of information are usually transferred alongside the main dataset to describe and codify the information or how it has been used and modified.

Many users may not be aware of the metadata being generated, collected, and shared with their documents and data, but missing metadata can cause significant problems for analysts.

“Typically, when you do machine learning and/or predictive analytics, you complete the process and then generate a report that explains the results of the analytics,” explained Jans Aasman, PhD, CEO of Franz Inc.

“But if someone wants to continue the research later or

replicate what happened the first time, you only have a paper document and you can't do anything programmatically with the results of your earlier analysis."

"So it is very important to generate and store metadata about the analytics that you did. Who did the analytics? What kind of scripts did they use? What techniques, and when? Which patients were involved? All of that metadata needs to be stored in the database as well."

Without this second layer of data detailing what has been done and why, it is difficult to repeat similar queries in the future, especially if a different analyst or informaticist is sitting down at the keyboard with no direct memory of what happened six months ago, added Parsa Mirhaji, MD, PhD, Associate Professor of Systems and Computational Biology and the Director of Clinical Research Informatics at the Albert Einstein College of Medicine and Montefiore Medical Center-Institute for Clinical Translational Research.

"One of the biggest problems in the big data environment is that people do all kinds of different analytics, and they may produce wonderful results, but it's often hard to tell if those results are accurate and meaningful," he said.

"Future analysts have to be able to find the information they need from previous reports, even when they don't know exactly what they're looking for. That is at the core of what we call the 'analytics tapestry.'"

Read: Top 4 Basics to Know about Semantic Computing in Healthcare

---

"You must be able to query data. It must be findable, systematically, without necessarily knowing beforehand if the

data exists. That is the only way to extract additional value from the data you have, and keep extracting that value over time."

Mirhaji and Aasman have been working together on the Semantic Data Lake for Healthcare at Montefiore Medical Center, a sophisticated machine learning project that is slated to go livefor patient care this summer.

The Data Lake will be used to provide predictive analytics capabilities to clinicians, starting with the ability to flag patients entering the hospital at a high risk of experiencing a serious crisis event within 48 hours. The system will also generate a customized checklist of intervention tasks that may help to avert or lessen the impact of the crisis.

Perhaps equally as important as the Data Lake's ability to deliver such advanced insights is its capacity to keep refining its results over time. The "learning" system will continue to make new connections across disparate categories of data to generate unique insights, but not without a comprehensive layer of metadata to weave the tapestry together.

"At Montefiore and Einstein, we are now capable of building these systems in a way that the metadata provides a framework for all different types of interactions between data sets," Mirhaji said. "If you are doing two different types of analytics on the same group of patients, this framework will help each algorithm 'find out' about each other automatically."

"Now if one project is working on predicting readmissions, and the other is working on risk stratification for mental illness, it certainly makes sense for each algorithm to pay attention to what the other one is doing, because there's a strong chance of interaction between those two areas of research."

With metadata to link these projects together without any extraordinary action needed by the analyst, it is possible to reduce unnecessarily repeated work and strengthen the results of each area of investigation, he added.

“So these many different small projects that are running in different corners of the system can come together and produce insights that were simply not available if everyone is working independently. And with the right metadata and machine learning algorithms, they can do that without any specific action by the researcher. That is how you build the core of the architecture that is able to expand its capabilities over time.”


Jans Aasman, PhD

Metadata also allows providers to utilize previous analytics for new applications, said Aasman. With a clear record of what tasks have already been performed, how they have been conducted, and what data has been used, it’s possible to tweak the methodology, add new datasets to the algorithm, or repeat the analytics on the same group of patients to gauge changes over time.

“You can try to find differences between the two analyses, or use the first as a benchmark for when you repeat the process a year later,” Aasman said.

“So if you are trying to do something like predict 30-day readmissions, you can see exactly what analytics you ran last year, and then use the same parameters a year later after making all sorts of workflow and process changes in the hospital to compare the results.”

The ability to track, record, repeat, and compare big data analytics work is increasingly important as the healthcare research community starts to share their activities on a

broader scale. As collaborations like the Precision Medicine Initiative and Cancer Moonshot begin to encourage an open ecosystem of research and data sharing, repeatability and reliability will only become more important.

“Research reproducibility is a foundation of science, yet one of the biggest issues in academia is the fact that researchers publish findings, make claims about the relationships between diseases and medications or interventions and outcomes, but no one can reproduce the findings because the publication contains no indication of what data was used, and there are no links to information about the analytics processes,” said Mirhaji.

Mirhaji and Aasman aren’t the only ones concerned that large-scale data sharing efforts may be putting integrity, quality, and accountability at risk. In a recent perspective piece for the *New England Journal of Medicine*, Laura Merson, Oumar Gaye, MD, PhD, and Philippe J. Guerin, MD, PhD also warned healthcare organizations against creating “data dumpsters” that have no usable metadata associated with them.

“There is currently inadequate funding and expertise for curating data to a standard and quality suitable for external secondary use,” the article said. “Researchers must bear the costs themselves or opt, as many currently do, to make raw data available without the explanatory documentation necessary to make them useful.”

“Most repositories are not equipped to rectify this problem — nor do they see this function as part of their mandate. As more partners in science mandate sharing of data, these platforms and repositories are likely to grow rapidly in number and size.”

Just like with a foundational EHR implementation, where taking some time to train users about data integrity at the beginning could save any number of clinical documentation headaches down

the line, paying close attention to curating metadata for the big data analytics environment is incredibly important.

“There are many, many different reasons why understanding each part of the analytics tapestry is very important,” Mirhai said. “It’s not just for governance, but for accountability and our need to reproduce findings in a reliable way. Those are all essential building blocks for the ability to create clinical programs from that data or make decisions for our patients.”

The Precision Medicine Initiative is one program that has attracted Montefiore Medical Center’s interest, and the organization’s competencies with semantic data analytics makes it a perfect participant.

“Precision medicine is an excellent application for these ideas,” said Mirhaji. “Montefiore was selected to participate in the White House’s inaugural Precision Medicine Initiative Summit because we are so interested in developing this, and because of our experience with the Semantic Data Lake, and efforts to combine data to customize patient care and tailor population health efforts to address community health needs.”

Read: How Precision Medicine Will Shift from Research to Clinical Care

---

The Data Lake already allows Montefiore to bring together disparate patient data sets, including clinical data, medical device data, and even socioeconomic information.

“That could also include genomics data,” Mirhaji pointed out. “All together, these data sets provide many more biomarkers on a real time basis than any one of these systems can provide individually. That’s what precision medicine is all about.”

“If you really want to solve these complex problems in healthcare and move away from the traditional way of treating patients based on a broad statistical average, you need a data system that goes beyond just data. You have to create a system that can link to anything outside your own predefined parameters – and that can learn from previous experiences, too.”

“That is where cognitive computing comes into the picture, and that is why you need to pay very close attention to how all of these data sets link together. Semantic computing is adaptable to those changes, if it is set on the right path to begin with, and that is what differentiates it from other forms of big data analytics.”