Knowledge Management World article – How does precision medicine become a reality? The Semantic Data Lake for Healthcare makes it possible

One of the prominent problems plaguing the current healthcare system is the narrow scope of patient data used to facilitate most aspects of care, from initial diagnoses to treatment. According to Dr. Parsa Mirhaji, director of clinical research informatics at Montefiore Health System and Albert Einstein College of Medicine, the vast majority of research findings are based on averages of middle-aged white males: “We don’t really know much about women, other ethnicities, children, you name it—there’s no evidence,” he says.

The White House launched The Precision Medicine Initiative in 2015 as a means of redressing the situation and expanding the breadth of patient data to personalize treatment for individuals and historically underrepresented groups. Achieving that objective requires not only amassing patient-specific data for wider demographics, but also storing, accessing and analyzing them with a number of avant-garde data management technologies including:

- the incorporation of a ***semantic data lake*** in which myriad information types such as billing codes, patient events, medical procedures and more are centralized and codified using semantic standards;
- the deployment of ***predictive analytics*** at a scale to leverage the aforementioned data to both anticipate and account for various patient outcomes in timeframes in which treatment can be administered to affect care;
- the use of ***machine learning*** algorithms to integrate the results of previous outcomes that significantly impact the analysis and effects of future patient objectives;
- the incorporation of an ***ontological pipeline*** that rapidly integrates new data, sources and requirements so data scientists can tailor models for highly targeted patient subsets; and
- the involvement of ***disposable data marts*** akin to sandboxes that quickly provision environments in which data scientists can manipulate data and analytics results while simultaneously allowing others to leverage that same data for their own purposes.

Thanks to the creation of the Semantic Data Lake for Healthcare (SDL), developed in partnership between Franz and Montefiore Medical Center, Mirhaji has been using those technologies and others to help achieve the Precision Medicine Initiative's goals. Initially, the researchers used a predictive analytics algorithm to flag any patient at risk of death or the need for intubation within 48 hours of being

hospitalized at any Montefiore Health Systems location. A personalized checklist of proposed interventions was then sent to the practitioners in charge of those cases.

Bolstered by the ability to determine nearly all the high-risk patients in the Montefiore population with approximately a one percent error, Mirhaji and his team have expanded the SDL's utility to hone in on personalized treatment options.

"Based on specific features, you put people in certain clusters," Mirhaji says. "Those specific features may dynamically change as we learn more about patients (e.g. new data about existing patients, new patients in the system). The algorithms identify optimal numbers of clusters and dynamically adopt best features to classify patients into those clusters. We don't make decisions about these things; this is all data-driven."

Data-driven semantics

The crux of the analytics deployed in the SDL depends on the semantic approach of the underlying infrastructure. According to Franz CEO Jans Aasman, the "maturity stack" powering the SDL centers around a distributed RDF graph database that turns everything in a traditional enterprise data warehouse into an exhaustive list of patient events. In addition, it involves a unified terminology system for healthcare, conduits for data science tools and more. It also includes what Aasman calls the "knowledge of humanities," which stems from various forms of external data. "There's a database called PubMed that has every scientific publication ever written," he explains. "People have turned that database into triples so that it can be easily linked into the SDL. There are gene ontologies that can be linked into the SDL platform. There are clinical trial databases, drug databases, disease databases. There's literally hundreds of available databases."

The RDF graph links that data in a highly contextualized manner due to its incorporation of semantic models that effectively standardize data regardless of structure. Those ontologies provide the foundation for the graph’s ability to determine relationships between specific data elements that users might not otherwise notice and are an essential component of the granular data integration required of such disparate sources. “This has been all driven by using semantics and ontological modeling behind it that enables robust data management and decision-making processes,” Mirhaji adds.