CMIO

Information, Evidence & Effectiveness in Medicine

Mining Clinical Data
Road to Discovery

Smartphones in Medicine
Dial 'M' for Murky

Radiology Images and MU
Vital, but Not in the EMR...Yet

The AMDIS Connection
The Ape Brain Meets the Medical Semantic Web

News & Views
› How to Stop Patient Data Breaches

Interoperability

GETTING BEYOND Point-to-Point Connections

contents

CMIO.net

cover story:

features:

departments:

web exclusives: CMIO.net

- Remote monitoring programs improve outcomes, save money
- Q&A: Adoption, not implementation, drives EMR success

June 2010

Mining Clinical Data
Road to Discovery

By Justine Cadet

Healthcare systems are utilizing new techniques to mine large amounts of biological, clinical and administrative data within clinical data repositories (CDRs) and EMRs. From these data, administrators and physicians are seeking out and often discovering a variety of previously unknown endpoints, including novel clinical associations for patients and quality indicators. These data discoveries can lead to better patient care and potential cost savings in the long run.

Practical Use

CDRs are increasingly becoming available as healthcare systems integrate patient information for research and utilization objectives. At the University of Virginia Health System in Charlottesville, William A. Knaus, MD, chair of the department of public health sciences, and colleagues conducted a proof-of-concept study to investigate the potential value of searching its CDR for novel insights, by applying a new data mining approach, IBM's HealthMiner, to a cohort of 667,000 inpatient and outpatient digital records.

"Today, most medical experiments are conducted with a hypothesis mapped out beforehand. With the mining tool, we discovered associations, and potentially new hypotheses that were plausible without having to train the system for these specific algorithms," Knaus says. With this wealth of data, he suggests medical researchers may discover previously unknown connections *a priori*.

For instance, Knaus et al observed a "strong correlation" between paralysis, peptic ulcer disease and renal failure, which was previously unreported; as well as a correlation between valvular disease, warfarin and cardiac arrhythmia, which has been extensively reported in clinical journals.

At the University of Michigan Health System, the Electronic Medical Record Search Engine (EMERSE) tool, which was developed in-house, searches the locally developed CDR that feeds into CareWeb, the locally developed EMR. The UMHS CDR contains more than 600 GB of free text data representing both inpatient and outpatient encounters for nearly 4 million distinct patients over the last 12 years. In 2009, new documents were being added to the CDR at a rate of 3.3 million per year which included 126 million lines of text.

According to David Hanauer, MD, assistant director of Comprehensive Cancer Center Bioinformatics Core at the University of Michigan (U-M) in Ann Arbor, EMERSE allows users to search for a list of terms or phrases, including the ability to support complex searches requiring wildcard matching or case sensitivity. While EMRSE requires a human interface for scientific deductions or conclusions, Hanauer says this method provides researchers with more accurate and complete data with more efficiency than the standard, manual medical chart review.

Hanauer et al conducted a similar study to Knaus, finding novel associations, such as those between granuloma annulare and osteoarthritis, and between pyloric stenosis and ventricular septal defect. They discovered that these coded data sets often don't contain details

that can only be found within the free text narrative reports that are best unlocked via natural language processiong (NLP) tools or EMERSE or some other automated/semi-automated mechanism, Hanauer says.

> "We are coming to an era where we can have some comfort basing our clinical decisions and recommendations on either clinical registries or data mining-based evidence, as opposed to an experimental approach."
>
> William A. Knaus, MD, Chair, Department of Public Health Sciences, University of Virginia Health System, Charlottesville

Human interfacing

At this point, data mining tools require some degree of human interaction because they cannot currently draw conclusions without human review. "With clinical mining, we're trying to surmount the research limitations of human beings attempting to query such large data sets," Knaus says.

Medical researchers are seeking to develop algorithms and methods to more accurately mine clinical data. "Up to this point, medicine has been a reductionist science," Knaus points out. "Clinical mining requires us to start at the top—at a population level—and figure what it means to go down to the data, as opposed to coming up from specific biological associations. It's a fundamentally different way of thinking for how we discover new connections and conduct medical research."

"By combining large amounts of data, you need this type of data mining or computational learning, along with increasingly sophisticated search-engine algorithms, to arrive at meaningful conclusions," says Knaus.

Human interaction also is required because of the incidental or false-positive findings that can occur when mining large data sets. "The false positives need to be manually separated from the clinically viable findings. The number of observations is much, much larger than the relevant data to be extracted," Knaus points out.

Challenges of interoperability, unstructured reporting

One major data mining challenge is integrating different biological and clinical data sets, says Knaus. Linking information from various data sources in the same clinical data set—for example, integrating phenotypic and genotypic data types—and getting them to talk with one another, is a big obstacle, he says.

With its data mining searches, U-M has experienced more difficulty with interoperability of the various databases than with different clinical data sets. "CareWeb tries to integrate data from all these disparate sources, so the users can view all data in one location, or at minimal, the most important documents," says Hanauer, who points out that the OB/GYN, emergency and surgery departments all use their own reporting systems.

However, Hanauer sees bigger problems related to data input and the lack of reporting structure. Most of the time, physicians employ free form text that is integrated in the EMR, and "free form makes it difficult to extract data computationally," he says.

"It's very difficult to get clinicians to enter coded data, as opposed to entering patient conditions. For instance, if a physician cannot easily access the code for strep throat, he or she will likely enter 'strep throat' or 'streptococcal infection' or 'strep A infection.' All that variability can lead to difficulties with data extraction," Hanauer says.

"For administrators and researchers, who need to extract data to develop reports, they would prefer all clinicians to enter information in the exact same manner, which means time is saved for the administrator, but not for the clinician," Hanauer says. "It's a tradeoff of who spends more time and who saves more time."

That structure has various facades, depending on the specialty. For instance, pathology has synoptic reporting, which allows for semi-structured reports for diagnosis or use-specific diagnostic codes, either SNOWMED or ICD-9 codes.

As a result, it is often difficult to seamlessly extract data. For example, a U-M surgeon involved with the National Surgical Quality Improvement Program conducted an EMERSE search of surgical site infections (SSIs). "In the documentation, no one had identified either surgical site infections or SSI. However, the reports mentioned pus in the wound or dehiscence of the wound—all of which may indicate SSIs, but the specific term was never mentioned," Hanauer explains.

NLP advocates are looking for ways to make the process as automated as possible without the need for human review of each case, according to Hanauer. For this to happen, a software tool would have to be trained to automatically review specific terms, and then an algorithm would have to be developed to determine which patients had SSIs. This NLP process would require a great deal of upfront expertise and time, he says.

The ability to produce quality enhancing, usable data depends on initial data entry, Hanauer says. "With EMERSE, if someone is searching for a quality indicator that isn't recorded in the EMR, they won't find the data," he says. "However, if a healthcare system is interested in a particular quality measure, the onus is on the administrators to ensure that the physicians are properly recording it."

Relevance amid healthcare reform

Knaus speaks to the role of clinical mining in the current era of healthcare reform. "The Obama Administration has placed great emphasis and funding on comparative effectiveness, in seeking to discover which treatments work for which patients. The only way this type of research will take place is to utilize large databases to make these comparisons," he says.

"We are coming to an era where we will have to develop some comfort basing our clinical decisions and recommendations on combinations of either clinical registries or data mining-based evidence, as opposed to data obtained from an experimental approach," Knaus concludes. CMIO