EMERSE: How Better Big Data Search Could Change Healthcare

January 16, 2019 | The electronic health record (EHR) has become both a blessing and a curse for many people in healthcare. While users are often frustrated with these systems, there is no denying that capturing all of the clinical data electronically is enabling new types of capabilities that were previously not possible with paper-based systems. Data within EHRs are often captured in free text clinical documents that provide many details about the patients—details that are not present in other parts of the system. These free text documents often remain an untapped treasure-trove of data because of the difficulty of using the information trapped within them. Software tools are now available that can help to leverage the data in these documents, one of which is EMERSE, the electronic medical record search engine.

Bio-IT World spoke with **David Hanauer**, physician and associate CMIO at the University of Michigan, about the differences between EMERSE and other natural language processing (NLP) systems, as well as the challenges of and predictions of EHRs more generally.

Editor's Note: Hanauer is speaking on EMERSE during the Bio-IT World West program this March 10-15, 2019, in San Francisco in the Bioinformatics for Big Data track.

Bio-IT World : What is EMERSE, and how is it similar or different from (NLP) tools?

David Hanauer: EMERSE is a search engine for free text clinical documents. Technically it's an information retrieval tool and not an NLP tool. EMERSE does have a lot of similarities with NLP but also some fundamental differences. At its core, EMERSE was designed to quickly identify clinical documents with the terms/concepts of interest, which could even be a single document among millions, and then display it to the user. NLP tools often have powerful features but can be difficult to setup and often have to be customized for each use case. This can mean involving NLP experts for each task or project. EMERSE is meant to be simple to use and doesn't require advanced training or an advanced degree to operate. Most people who can use Google can also use EMERSE. EMERSE does not automatically code data; rather, it helps users identify data quickly and leaves it up to the user and their intelligence to interpret the meaning in the context of the documents. NLP tools often automatically code the data, but getting the context and nuances correct is challenging, which can lead to errors that are difficult to detect. NLP tools certainly have their place in a clinical or research environment, and I think EMERSE does too. In many ways these tools provide complementary functionally.

How does EMERSE differ from a typical search engine?

EMERSE was built specifically to help with searching clinical documents, so while it is not an electronic health record (EHR) system, it does provide features that would be expected of an EHR. For example, documents are grouped by patient, and searches across the entire set documents will return counts based on the number of unique patients and not unique documents. This kind of feature is needed to estimate cohort sizes, among other tasks. A typical search engine would not group documents together by patient. EMERSE also supports the use of query expansion with hundreds of thousands of terms including trade and generic drug names, medical acronyms, and abbreviations, and other wording variations and misspellings that often appear in the clinical documents. However, rather than automatically expanding all query terms, EMERSE will present the options to users to let them pick from just the ones they want. EMERSE also provides the capability to make a search case sensitive when needed. For example, to distinguish the common word 'all' from the medical acronym 'ALL' which stands for acute lymphoblastic leukemia.

What are some use cases for a medical record search engine?

The use cases are broad and varied. At a high level, searching clinical documents supports both the operational and research enterprises. For operations it can be used to help groups as diverse as infection control, health information management, and coding compliance teams. For research it aids in cohort identification, eligibility determination, data abstraction, and more. In fact, EMERSE has been used in over 1,500 clinical research studies and on the operational side it is both a time and money saver. We have found that most users can't go back to their manual processes once they have experienced the efficiencies of a medical record search engine.

Where is EMERSE being used and how can people learn more?

EMERSE has been used at the University of Michigan for over a decade, and is now also being used at the University of North Carolina. With support from the National Cancer Institute, we have plans to implement EMERSE at other sites including Case Western Reserve University, the University of Cincinnati, Columbia University, and the University of Kentucky. EMERSE is available at no cost and is open-source. Many details about the system can be found on our website at http://project-emerse.org

.

What are some of the challenges in extracting data from an EHR?

There are a variety of issues that make extracting and utilizing data from an EHR difficult. First, few people are ever allowed to access the data. This is partly due to security and compliance reasons, partly to ensure patient privacy, and partly because there may be concerns about accessing data from a production system that could be slowed down by other systems accessing the same repository. Some organizations have built reporting databases, and some EHR vendors include such reporting databases as part of their standard build. Even in those cases it can still be difficult. For example, we've worked with a common reporting database where all of the document formatting is stripped before being added to the database. That can make the text easier to work with in some circumstances, but it makes it less valuable when presenting the text to a user on the screen since elements such as table structures and other formatting becomes lost.

How do you anticipate EHRs involving in the next 10 years?

Predictions are always difficult, but I would guess that market consolidation among a few large vendors will continue. Innovation in the EHR space seems to be slower than other domains, but I believe EHRs will continue to slowly add new features and functionality. It is not clear if vendors will be able to make each component as feature rich and easy to use as more specialized, standalone systems, the latter of which can often be more nimble in their development and less constrained by legacy code and monolithic user interfaces. I hope that EHRs continue to "open" up more to allow more applications to be built around or on top of them, leveraging the EHR's database as a source of truth but allowing for more innovation on the periphery. Standards like the Fast Healthcare Interoperability Resources (FHIR) will likely help to spur this kind of innovation. Apple has demonstrated this type of innovation around EHRs with their Health App on the phone and ECG app on the watch. Integrating these types of external data into a patient's medical record will likely occur over time, but vendors will have to be careful about not causing information overload for clinicians with all of these extra data. EHRs will likely provide more overall functionality to engage the patients and involve them in their own care through portals or other approaches, again possibly using standards like FHIR through personal devices. EHRs will also likely start to include more predictive modeling and AI to help clinicians find patterns or make better predictions and decisions with the growing body of data stored in the EHR. Nonetheless, a big hope of mine is that EHR vendors will take a break from adding features and focus more on making the systems more user friendly and intuitive.

Will machine learning and artificial intelligence (AI) eventually replace the need for software such as a search engine?

Maybe someday, but probably not in the near-term. Deep understanding of clinical text has been more challenging than many people have assumed. It is worth pointing out that medical text is often far from "natural" language. The documents often have many ambiguous abbreviations, conflicting information, and other aspects that make accurate interpretation difficult not only by a computer but even a human. I think even the IBM Watson team has acknowledged that. These are all very active topics of research in the world of machine learning and NLP. Augmenting human intelligence with systems that can present information to a person efficiently is still a very valid and workable approach, even if the person has to ultimately make a decision about the coding of the data elements of interest.

View Next Related Story
How Common Data Could Lead To Uncommon Alzheimer's Discoveries | Jan 14, 2019

CHI
A division of Cambridge Innovation Institute (CII)
250 First Avenue, Suite 300
Needham, MA 02494

P: 781.972.5400
F: 781.972.5425
E: chi@healthtech.com