ICBO: International Conference on Biomedical Ontologies · Buffalo, NY, USA
Doctoral & Post-Doctoral Consortium, July 27, 2011

The Age of Data-Driven Medicine: Mining the Electronic Health Record

Paea LePendu, Mark A. Musen, Nigam H. Shah

Stanford Center for Biomedical Informatics Research, Stanford, CA, USA

With the availability of tools for automated coding of unstructured text using natural language processing, the existence of over 250 biomedical ontologies, and the increasing access to large volumes of electronic medical data, it is possible to apply data-mining techniques to the large amounts of unstructured data available in medicine and health care. For example, by computationally encoding the free-text narrative—comprising roughly 80% of the clinical electronic medical data—it may be possible to test drug safety signals in an active manner. We describe the application of NCBO Annotation tools to process clinical text and the mining of the resulting annotations to compute the odds ratio of having a myocardial infarction on taking Vioxx for Rheumatoid arthritis. Our preliminary results demonstrate that it is possible to apply annotation analysis methods for testing hypotheses about drug safety using electronic medical records.

Given recent advances in detecting drug safety signals from spontaneous reporting systems, it becomes crucial to develop methods of searching for, testing, and applying these signals throughout the electronic health record so as to realize their benefits on new patients before an adverse event occurs. We hypothesize that using ontology based approaches, analogous to enrichment analysis, can fill this gap.

One of our recent successes has been the considerable improvement of the NCBO Annotation tools. We have optimized the system for both speed and space so that it can extract clinical concepts from the textual reports of the entire Stanford Clinical Data Warehouse—nearly 10 million surgical pathology, radiology, and general transcription reports—overnight, on a single machine. Furthermore, we have added negation detection as well as extended concept recognition capabilities, such as morpheme-based matching. We are highly encouraged by our preliminary results in detecting the Vioxx risk signal from that data.