Extracting Information from Electronic Health Records Using Natural Language Processing – Knowledge Discovery from Unstructured Information

Vasu Chandrasekaran[1], Jinghua He[1], Monica Reed Chase[1], Aman Bhandari[1], Christopher Frederick[2], Paul Dexter[2]
[1]Merck & Co., Inc., [2]Regenstrief Institute

Background. Extracting meaningful data from electronic health records in an accurate and efficient manner is challenging, with a significant portion of clinical information found only in unstructured, free-text clinical documents. Critical information captured by physicians is rarely available in coded format, which limits the opportunities for clinical decision support or quality monitoring. Even a frequently used metric such as the - ankle-brachial index (ABI) – a 'quantitative' data point for defining peripheral arterial disease (PAD) is typically embedded in the text of radiology reports and not found in structured datasets. Natural Language Processing (NLP) methods provide an automated solution for extraction of information from clinical text and could substantially reduce the burden of manual reviews which can be time-consuming, costly and error prone.

Methods. nDepth™ is a text-mining solution developed at the Regenstrief Institute for the development and validation of NLP algorithms across one of the largest health information exchanges in the Nation, the Indiana Network for Patient Care. It uses a combination of linguistics, pattern recognition and machine learning to derive meaning from narrative clinical text. The solution uses a SOLR index for sub-second retrieval of free text data from clinical notes, extracts context-appropriate information from reports and processes complex search queries to return meaningful results. A highly effective component for value extraction is the state machine, which is a method to create and execute a series of regular expressions one after another, deciding which one to execute next based on the result of the execution. For example, string "A" identified the name of a drug and string "B", identified the dosage. The state machine would then extract the values paired the result of the string A with the result of string B, identifying the dosage of drug 'A' was 'B'.

Results. The NLP-based text-mining platform has been refined through extensive and repeated use searching over 230 million text records from more than 17 million patients. It has been extensively validated through research performed in collaboration with scientists at Merck, including 1) the identification and determination of disease severity for conditions such as heart failure and malignancy, as well as findings from radiology or pathology reports, 2) extraction and conversion of unstructured quantitative data to structured observations (such as ejection fraction, MMSE, Karnofsky score), 3) assessment of patient characteristics such as emotional and social behaviors.

We recently demonstrated that the identification of patients with peripheral arterial disease in observational data was significantly increased employing NLP. We identified four-fold more patients using a conservative (i.e., high specificity) NLP algorithms than by using an established code- based strategy for PAD detection. Overall 43,811 unique PAD patients were identified across all methods. NLP identified 95.2% compared with 21.9% for ICD9/CPT codes and 9.9% for ABIs alone. Over 75% of patients with physician-documented evidence of peripheral arterial disease were not identifiable by structured data.