Obtaining Phenotypic Data in Oncology via Natural Language Processing and Data Mining of Text Topics

Rebecca A. Ottesen
City of Hope, Duarte, CA USA – rottesen@coh.org

Courtney A. Vito
City of Hope, Duarte, CA USA – cvito@coh.org

Joyce C. Niland*
City of Hope, Duarte, CA USA – jniland@coh.org

While complication rates after lumpectomy or mastectomy in breast cancer patients are usually low, when complications do occur they lead to increased patient morbidity, extended length of stay, additional procedures, and higher cost. We applied data mining techniques to interpret phenotypic data and discover patterns that are potentially predictive of patient outcomes and surgical complications. Text topics were created using International Classification of Diseases (ICD-9) codes, as recorded in the medical record at the time of surgery. We also employed Natural Language Processing (NLP) of unstructured medical dictations to enhance the codified data that was available in our integrated enterprise data warehouse. Groupings of diagnostic and procedure ICD-9 codes were assembled using text topic and document cluster modelling through singular value decomposition techniques. Several data mining tools such as decision trees, neural networks, and logistic regression were evaluated to assess the predictive value of the topic groupings. The goal of our analysis was to help identify patients at higher risk who may require pre-emptive intervention perioperatively. This talk will describe our cohort, methods, analysis, results and implications of applying NLP and data mining on patient information to predict complications of surgery in an attempt to improve patient outcomes.

Keywords: data mining; text topics; natural language processing; healthcare research