DEDUCE Clinical Text: An Ontology-based Module to Support Self-Service Clinical Notes Exploration and Cohort Development

Christopher Roth, MD, MMCi, Shelley A. Rusincovitch, Monica M. Horvath, PhD, Stephanie Brinson, Steve Evans, Howard C. Shang, Jeffrey M. Ferranti, MD, MS

Duke Medicine, Durham, North Carolina

Abstract
Large amounts of information, as well as opportunities for informing research, education, and operations, are contained within clinical text such as radiology reports and pathology reports. However, this content is less accessible and harder to leverage than structured, discrete data. We report on an extension to the Duke Enterprise Data Unified Content Explorer (DEDUCE), a self-service query tool developed to provide clinicians and researchers with access to data within the Duke Medicine Enterprise Data Warehouse (EDW). The DEDUCE Clinical Text module supports ontology-based text searching, enhanced filtering capabilities based on document attributes, and integration of clinical text with structured data and cohort development. The module is implemented with open-source tools extensible to other institutions, including a Java-based search engine (Apache Solr) with complementary full-text indexing library (Lucene) employed with a negation engine (NegEx) modified by clinical users to include to local domain-specific negation phrases.

Introduction
DEDUCE is a Web-based platform that allows clinicians and researchers to securely access data from the EDW in a self-service model. Enhanced support of clinical text was a logical extension for this informatics platform. The new clinical text functionality is tightly integrated into the suite of tools available within DEDUCE, including the cohort management tools and data export mechanisms, and allows for seamless transitions between system modules.

Background
We deployed two ontologies to support the new module. The extensive SNOMED CT ontology was selected because of its growing use in health policy. RadLex, a domain-specific ontological lexicon supported by large radiological professional organizations, was also chosen specifically to support radiology reports. The extensible ontology framework of DEDUCE Clinical Text can support any ontology that conforms to a predictable hierarchical structure. The search capabilities leverage ontological relationships, so that conditions, synonyms and “child” terms are presented to the user when searching and filtering reports.

Methods
The system is implemented through several complementary mechanisms that use text tokenization, sentence splitting, and word stemming. Nightly batch jobs process new documents (using predefined rules for different document types) to identify document sections, annotate and tag ontological terms, and parse text through the negation engine. When a user submits a text search, which can include advanced parameters such as term proximity and fuzzy matching, the query string is analyzed in real time and passed to the search engine. The returned corpus of documents profiles all search terms, outlines ontological relationships, and visibly highlights for the user search terms or phrases in the report output. The integration with other DEDUCE modules allows cohort exploration to be interactive with text search results, and allows clinical text to be exported with structured data. The system is implemented using a Java-based search engine (Apache Solr) with complementary full-text indexing library (Lucene) employed with a negation engine (NegEx) modified by clinical users to include to local domain-specific negation phrases.

Results and Discussion
The DEDUCE Clinical Text module was first deployed in September 2011 with radiology reports and later expanded to include pathology reports. Case studies from our user community demonstrate that the new functionality is valuable for its integration into research cohort development, medical education, operational optimization, and clinical care improvement. The open-source tools used for system implementation can be utilized by other institutions and provide an extensible environment for future development. The module provides a robust framework, and is scalable for incorporating new sources of clinical text and ontologies.