# Extracting Relevant Information from Clinical Records: Towards Modeling the Evolution of Intracranial Aneurysms

**Juan Anna Wu[1,2], BSc, William Hsu, PhD[1], Alex A.T. Bui, PhD[1]**
**[1]Medical Imaging Informatics Group, Department of Radiological Sciences**
**[2]Biomedical Engineering Interdepartmental Program**
**University of California, Los Angeles, CA**

**Abstract**

*We present an open-source natural language processing tool that reduces the amount of manual labor required to extract pertinent information from clinical documents and populate an existing data model. The system is built upon the Unstructured Information Management Application (UIMA) framework, which provides a flexible pipeline for analyzing and annotating medical records. Our initial evaluation showed a precision and recall of 96.8% and 64.8%, respectively. We apply this tool towards studying patients with intracranial aneurysms.*

**Introduction:** Intracranial aneurysms (ICA) are pathological enlargements of cerebral arteries that are often associated with high morbidity and mortality rate when ruptured. Understanding the origin, progress, and treatment of this disease is important for creating a model that predicts risk of rupture. However, much of the information is locked within free-text clinical reports, requiring manual extraction by a domain expert. The goal of this project is to automatically extract relevant information, such as morphology and temporal information from medical record.

**Methodology:** First, key findings and attributes of ICAs are modeled by reviewing published literature, interviewing physicians, referencing other models such as @neurist and manually examining patient records. The data model includes variables representing clinical, imaging, histopathology, and genetic information. We performed an initial study to determine which variables could be consistently extracted from medical records based on how they are reported and what documents contained the information. In this initial effort, we have targeted 94 different variables which include ones related to patient demographics, medical history, social and family history, clinical presentation and outcome, morphology, imaging follow-ups, treatment, and hospital course. This information is reported primarily in admission and discharge summaries, radiology and surgical reports, referral letters, and neurology consultation notes. Building upon Unstructured Information Management Architecture (UIMA) framework [2], we implemented pipeline of annotators to target the aforementioned variables: 1) a regular expression annotator extracts variables with consistent representations (e.g., aneurysm neck measurement); 2) a part-of-speech tagger is used to identify noun phrases; 3) a dictionary lookup method is used to map noun phrases to biomedical ontologies such as Foundational Model of Anatomy (e.g., to extract anatomical locations of the aneurysm); 4) a status annotator based on ConText [3] is used to identify negation and temporality (history, recent, or hypothetical) of a finding. The extracted results were inputted into our data model, represented as a relational database.

**Evaluation:** Evaluation was conducted on 398 documents from 20 patients seen at the UCLA Medical Center. Medical fellows were asked to manually review each patient case to generate a gold standard. We achieved a precision of 96.8% (among 838 fields filled, 811 are correct.), and recall of 64.8% (among 1252 fields filled, only 811 field are correct).

**Conclusion:** Our preliminary result shows that the UIMA framework with custom annotators can be used to populate a data model with information from the patient record. Currently, the system is being deployed to assist the physicians with manual data entry; manual oversight is still needed to ensure the quality of the inputted data. Future work includes: 1) performing co-reference resolution to map semantically related variables together and 2) generating a patient-centered visualization and predictive model using the data extracted by our system.

**References**

[1] Carol Friedman. Semantic text parsing for patient records. In H. Chen, S. Fuller, C. Friedman, and W. Hersh; Medical Informatics, editors, USA: Springer. 2005.

[2] David Ferrucci and Adam Lally. 2004. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

[3] Chapman,W, Chu D, Dowling JN. (2007) ConText: An algorithm for identifying contextual features from clinical text. BioNLP 2007: Biological, translational, and clinical language processing, pp. 81–88.