Preliminary findings in natural language processing to stratify patients with mental illness

Myung Woo, MD[1], Dylan Liu, Stephen Evans, Jane P. Gagliardi, MD, MHS[1,2], Jessica D. Tenenbaum, Ph.D.[3]
Departments of Medicine[1], Psychiatry[2], and Biostatistics & Bioinformatics[3], Duke University, Durham, NC
✉ Jessie.Tenenbaum@duke.edu

Abstract

Comptutational phenotyping has become an important tool in the Learning Healthcare arsenal in order to identify patients with a given condition or phenotype. Though diagnosis codes, labs, and medications may be helpful for cohort identification, natural language processing (NLP) of clinical notes can help improve accuracy. For mental health disorders in particular, which tend to be highly heterogeneous and overlapping, NLP can help define more homogeneous cohorts.

We used cTAKES, an open source Apache standard software package for NLP of clinical notes, to extract biomedical terms derived from SNOMED CT, from patient records at Duke. Despite cTAKES' optimization for clinical text, acronyms proved to be a significant challenge, both in false positives and false negatives. Preliminary results demonstrate the importance of domain knowledge and sanity checking of black-box big data approaches.

Background

Mental health disorders tend to be defined using an "à la carte" menu approach in which a person is diagnosed with a given disorder if he or she demonstrates, e.g. 2 out of a list of 5 possible symptoms. As a result, two different patients with completely non-overlapping symptoms may both be diagnosed with the same disorder, e.g. schizophrenia. Though this heterogeneity is acceptable for epidemiological tracking and even for treatment decisions, it makes interrogation of a biological basis for disease much more challenging. In order to leverage the Learning Health System for research into disease etiology and mechanisms, it is necessary to be able to identify more homogeneous cohorts based on actual symptoms. This work represents a first foray into patient stratification in schizophrenia using NLP applied to clinical notes.

Approach

cTAKES™

clinical Text Analysis Knowledge Extraction System

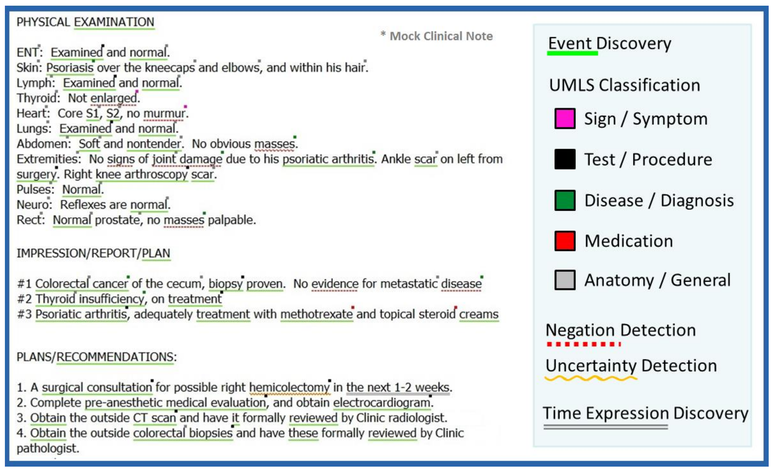

Figure 1: cTAKES functionality. cTAKES reads in a clinical note and parses text to identify a specific set of semantic concepts as well as their context, including negation, uncertainty, and expressions relating to time. The set of semantic concepts may be from an existing terminology, or a custom built dictionary, or both. (Note that this example is from http://ctakes.apache.org/whycTAKES.html and does not use Duke data nor our customized dictionary.)

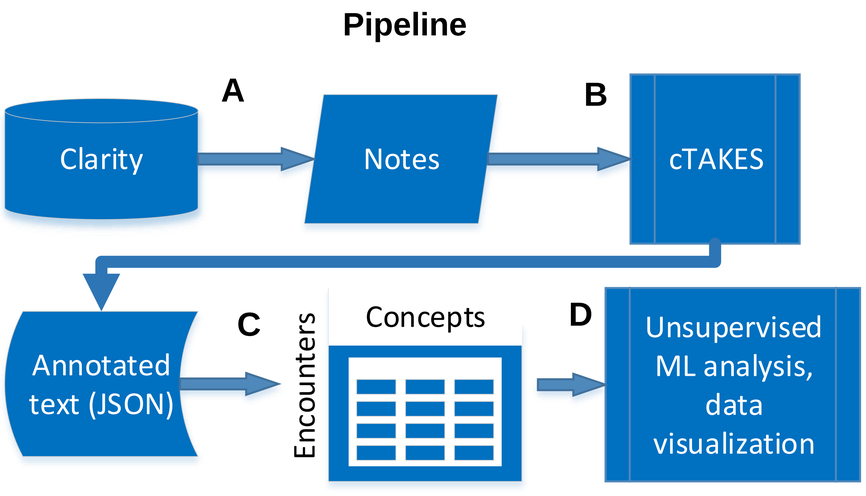

Figure 3: NLP Analysis Pipeline
A. Notes were extracted from Clarity and stored as text in a GreenPlum database. **B.** Each note was run through cTAKES, which used a specific semantic subset of SNOMED CT as well as a custom built dictionary primarily for acronyms, generating a JSON file that indicated actual observed text, extracted term, semantic type, and negation status. **C.** For each note, terms were evaluated for frequency and negation. Encounters with multiple notes were consolidated into a single value indicating presence and negation. **D.** Sparse data was clustered using Latent Dirichlet Allocation (LDA), a Bayesian statistical model widely used for topic modeling in NLP and visualized using tSNE (T-distributed Stochastic Neighbor Embedding).

Biomedical Terms

SNOMED CT
The Systematized Nomenclature of Medicine (SNOMED) is a systematic, computer processable set of terms from the biomedical domain. Each term has one ore more semantic types, such as symptom, mental process, mental or behavioral dysfunction, etc.

Custom Dictionary
Though SNOMED includes concept synonyms and some acronyms, a number of acronyms commonly used in psychiatric notes are not included. To address this gap, we created a custom dictionary to detect frequently used terms such as AVH - audio-visual hallucinations; PI- paranoid ideation; SHI - suicidal/homicidal ideation.

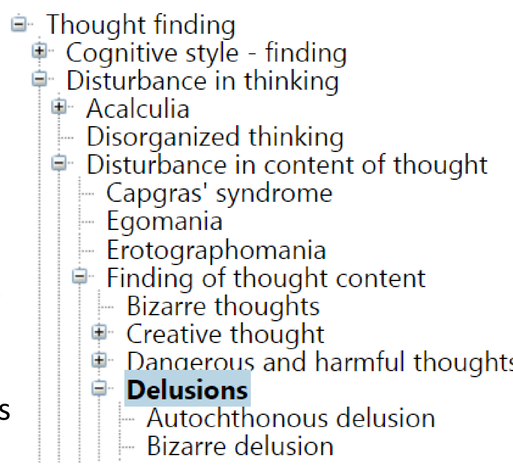

Figure 2: SNOMED subset

Results & Conclusions

Preliminary analysis of extracted concepts indicates areas for iteration in application of cTAKES before useful signal may be extracted. Some examples are:

- A large number of false positives resulted from a common word being interpreted as an acronym, instead. For instance, "Plan" was designated as "PLA2G6-Associated Neurodegeneration," which is a synonym for "Infantile Neuroaxonal Dystrophy," which is not what the documenting provider intended by outlining a *plan* of care.
- Some synonyms are being mapped to inappropriately specific subterms. E.g. the term "sensitivity" is interpreted as a mood finding, but also mapped by default to "antimicrobial susceptibility."

Next steps will be to update specific parameters for cTAKES to address items above and generate and refine results to optimize for diagnostic subgroup stratification.

Acknowledgements

The authors would like to acknowledge Bass Connections; team members Scarlett Zhou, Abhi Jadhav, Sanya Kochhar, Aakash Thumaty, Kamyar Yazdani, and Casey Riffel; Allison Young, Dr. Gopalkumar Rakesh, Dr. Nigam Shah, Dr. John Beyer, and Colette Blach. JDT was funded by grant K01 LM012529 from the National Library of Medicine.