NLP Working Group

Chunnan Hsu
Biomedical Informatics
UC San Diego
October 13, 2016
Current members

- Chunnan Hsu
- Ramana Seerapu
- Scott Duvall
- Olga Patterson
- Hua Xu
- Michael Matheny
- Glenn Gobbel
- Tsung-Ting Kuo
The NLP working group is tasked to accurately extract phenotypes for three clinical conditions: Kawasaki Disease (KD), Weight Management / Obesity (WM/O), and Congestive Heart Failure (CHF), from tens of millions of clinical notes shared by participating institutes in pSCANNER, and seamlessly integrate with the shared structured data.
Milestones

- Refine the problem definitions of the extraction of common data elements to guide the finalization of a common evaluation guideline, mapping to a common output data model, and design, reuse, and sharing of NLP tools.
- Finalize output format definitions with OHDSI.
- Create secure, privacy-preserving, cross-institution clinical NLP infrastructure where tens of millions of clinical notes can be processed and the quality of processing can be assessed semi-automatically.
- With the infrastructure, create large data warehouse of NLP extracted data from clinical notes to support phenotyping of the three pSCANNEER use case conditions.
Workflow

Cohort Identification from Clinical Texts (CICT)

IRB Approval

Target Phenotype

EHR Data Warehouse

Clinical Notes

iDASH Midas Interface

iDASH Midas

NLP Ensemble Pipeline

Common Data Elements

Annotated Clinical Notes

Annotation Review Interface (Remote Desktop)

Reviewed Annotations

Annotation Review Interface (Remote Desktop)

OMOP Common Data Model

Reviewed Annotations

Cohort Identification Interface

Phenotype Database

Phenotype Database

Retrieved Patient ID List

P002
P293
P534

Retrieved Note ID List

N112
N431
N886

UCSD AD/VPN
iDASH VPN Pool
iDASH 2-Factor Authentication
iDASH VDI Remote Desktop

Supported by the Patient-Centered Outcomes Research Institute (PCORI) Contract CDRN-1306-04819
Current Progress

EHR Data Warehouse ➔ iDASH Midas Interface ➔ iDASH Midas ➔ NLP Ensemble Pipeline ➔ Annotated Clinical Notes

Common Data Elements

UCSD AD/VPN
iDASH VPN Pool
iDASH 2-Factor Authentication
iDASH VDI Remote Desktop

Cohort Identification from Clinical Texts

IRB Approval

Clinical Notes

Reviewed Annotations

Annotation Review Interface (Remote Desktop)

iDASH Midas Interface

iDASH Midas

Common Data Elements

Annotated Clinical Notes
CICT: NLP Ensemble Pipeline

De-identified Clinical Notes

Common Data Elements

NLP Preprocessor
- De-identification
- Encoding Converter
- Sentence Splitter

NLP Toolkit
- cTAKES
- MetaMap
- CLAMP
- EFex

NLP Ensemble
- Intersection (Good)
- Union (Iffy)

NLP Postprocessor
- Annotation Tags
- Extracted Data Elements

Annotated Clinical Notes

Common Data Model
Cohort Identification from Clinical Texts

http://textmining.ucsd.edu:5005