Extracting OMOP Concepts from Clinical Narratives to Empower Clinical Research

January 25, 2022

https://covid.cd2h.org/
Outline

● N3C Overview
● Open Health Natural Language Processing (OHNLP)
  ○ Ethical AI framework
  ○ Scientific rigor
  ○ OHNLP Toolkits
  ○ PASC NLP algorithms
● Join the journey for NLP-empowered RECOVER
  ○ Contribute to the development effort
  ○ Implement the N3C NLP Infrastructure
  ○ Augment data with NLP for N3C
  ○ Collaboration and Partnership
N3C Enclave: The largest public HIPAA-limited longitudinal-EHR data set in US history
(1/20/22 release)

Hub Sites: 69
Persons: 10.7 million
COVID+: 3.8 million
Rows Data: 12.0 billion
Rows/Patient: ~1,120

https://covid.cd2h.org/dashboard/
Row level data:
- Person
- Drug
- Procedure
- Condition (diagnoses)
- Measurement (labs)

Types of encounters:
- Inpatient, ED & Outpatient
- Longitudinal back to 2018
Each of the 69 sites has a pipeline with 2M+ transformations

The provenance between 5000 syntactic transformations across the 69 sites is automatically tracked.

This enables:

● pipeline developers to very quickly identify the root cause of data quality issues

● data pipelines can be refreshed in <20 minutes whenever the source data updates
Open Health Natural Language Processing

Hongfang Liu
Mayo Clinic and IBM to move beyond EMR’s to deliver knowledge at the point of care

Posted at 07:26h in Health Technology, IT by WTN News • 0 Comments

Madison – Biomedical informatics researchers at Mayo Clinic and IBM have launched a Web site for the newly founded Open Health Natural Language Processing (NLP) Consortium. The consortium is establishing the open-source space to promote past and current development efforts, including participation in information extraction from electronic medical records. Mayo Clinic and IBM Healthcare released clinical NLP technologies into the public domain. The site will allow the approximately 2,000 researchers and developers working on clinical language systems worldwide to contribute code and further develop the systems. Additionally, the VA Boston Healthcare System and Seattle Group Health have strongly indicated their support of the concept according to IBM.

“We are inviting our international colleagues to help continue development of these valuable tools,” says Christopher Chute, M.D., Dr.P.H., Mayo Clinic bioinformatics expert and senior consultant on the project. “By making it an open-source initiative, we hope to enable wide use of these NLP tools so medical advancements can happen faster and more efficiently.”

**cTAKES** functionality recognizes whether a clinical concept is negated, relevant to the patient or to the patient’s family, which are attributes critical to understanding patient-centered medical care.

**IBM’s medKAT systems** (medical Knowledge Analysis Tool) is a system to extract structured information from unstructured data sources, such as pathology reports, clinical notes, discharge summaries and medical research publications.
Strategic Health IT Advanced Research Projects (SHARP)
Research Focus Area 4 - Secondary Use of EHR Data
Increasing efficiency of patient care through electronic healthcare records
● Research on implementability of open source NLP systems (https://github.com/nlpie/nlp-adapt) (led by Dr. Serguei Pakhomov)

● Research on usability to deliver production-ready NLP (led by Hua Xu)

● Research on NLP utility and generalizability for clinical research and practice (led by Hongfang Liu)
Aim 1: **Obtain** NLP artifacts across sites for improving NLP algorithm development

Aim 2: **Generate** a synthetic text corpus for exploratory analysis of clinical narratives

Aim 3: **Develop** privacy-preserving computational phenotyping enhanced with NLP

Aim 4: **Partner** with diverse communities for translation science excellence leveraging EHR
Human-centric Ethical AI Framework

FAIR Data Principles
- Findability
- Interoperability
- Accessibility
- Reusability

RITE Implementation Principles
- Transparent and Implementable: Process Correctness
- Explainable and Reproducible: Result Correctness
The Use of AI for Health
Guiding Principles

- Protecting human autonomy
- Promoting human well-being and safety and the public interest
- Ensuring transparency, explainability and intelligibility
- Fostering responsibility and accountability
- Ensuring inclusiveness and equity
- Promoting AI that is responsive and sustainable
Human-centric Ethical AI Framework

- Language plays an unique role in humanity and society.
- Need to follow a social and ethical framework when working on health language.

Too many AI researchers think real-world problems are not relevant

The community’s hyperfocus on novel methods ignores what’s really important.

by Hannah Kerner August 18, 2020
About 4,240,000,000 results (0.80 seconds)

The concept behind “data is the new oil” is that just like oil, raw data isn’t valuable in and of itself, but, rather, the value is created when it is gathered completely and accurately, connected to other relevant data, and done so in a timely manner. ... COVID-19 related data is being generated quickly.  Apr 27, 2020

https://www.kenwayconsulting.com › blog › data-is-the-n...  

Is Data Really “The New Oil”? | Kenway Consulting
Reproducibility and Methodology Rigor

- How is NLP evaluation conducted in clinical research?
- What’s the current reporting practice of the NLP component in the traditional clinical research?
- What is the degree of granularity when reporting NLP methodology and evaluation?
- What are the potential barriers to adopt clinical NLP to clinical research?
- What are the lessons learned from the existing evaluation and reporting practices
We observed high heterogeneity in the reporting practice: 14% of studies did not report NLP methodology and evaluation, 22% of studies did not report evaluation design, and 10% of studies did not report NLP methodology. A few studies claimed that ‘NLP has been evaluated prior to release’. However, no additional details can be found to justify the validity of the NLP results and clinical findings.
Reproducibility and Methodology Rigor

Evaluation Challenges

- Only mention cohort duration but not NLP evaluation duration
- EHR system migration, change of definition, ETL process
- Study setting and EHR environment
- Definition variation between cohort and disease
- Commercial vs open source
- Evaluation performance/matrix

Implication to Clinical Research

- Most data elements identified by NLP are either comprised of risk factors (48%) or outcome (42%), which have strong implication to the outcome of the study.
- The validity of the study is thus dependent on the rigor of NLP methodology and evaluation.
A Collaborative Framework for NLP Development

- N3C Domains
  - COVID-19 Signs and Symptoms
  - Long COVID
  - Aging
  - Adverse Event

The framework includes:
- Model Development Interface
  - Web GUI
  - Model/Algorithm
    - Knowledge-Driven
    - Expert-Driven
    - Corpus/Data-Driven

- OHNLP Backbone
  - OHNLP MedTagger
    - Concept Extraction - Contextual: Temporal, Subject, Negation
  - Dictionary Lookup
  - Other NLP systems

- Benchmarking

- Institutional Data Center
  - Local instance with NOTE and NOTE_NLP tables

- OHDSI

- N3C Enclave

Work in progress

OHNLP Toolkits
A Federated Framework for NLP Evaluation
A Minimal Viable Product (MVP) for NLP Deployment

Ingestion Layer
- RDBMS
- File System
- Option 1 N3C NLP Solution

Processing Layer
- N3C NLP Solution
- Textual Contents
  - Option 1 N3C NLP Solution
  - Option 2 In-House NLP Solution
- Named Entity Extraction
  - UMLS, MONDO, etc.
  - COVID-19 Concepts
- Ontology Dictionary Lookup
- Keywords/Regex
- Temporal Information
- Document Identifiers
  - Date/Time
- NLP Events
  - Rule Context

Persistence Layer
- OHNLP Toolkits
- Local

Retain
- OHDSI
- N3C Central
- NOTE NLP Table

Extract
- Run
- Populate

Optional 1 To Do
● Incorporate dictionary lookup for concept mentions
  ○ Default config includes COVID/ PASC relevant concepts
  ○ Dictionary resource for comprehensive NLP is available by request
    ■ A curated dictionary, MedLex, with OMOP Concept Identifiers mapped available
● Work on documentation for onboarding
Towards a semantic lexicon for clinical natural language processing

Hongfang Liu, Stephen T Wu, Dingcheng Li, Siddhartha Jonnalagadda, Sunghwan Sohn, Kavishwar Waghokar, Peter J Haug, Stanley M Huff, Christopher G Chute

Affiliations + expand

PMID: 23304329  PMCID: PMC3540492  Paperpile

Free PMC article

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540492/
PASC Resources


Dictionary Resource Generation

- Map MedLex to the concept identifiers from the latest OMOP Vocabulary (through standard vocabulary)
- All terms from the two PASC articles were processed and mapped to the concept identifiers from the latest OMOP Vocabulary.
- For PASC terms could not be mapped, we manually reviewed and aligned with the closest OMOP concept identifiers if available or map to the HPO identifiers.
Refinement with 97 long hauler notes
  ○ Add 23 concepts not included in the initial PASC dictionary but mappable to OMOP or HPO Vocabulary
  ○ Add 4 concepts not mappable to OMOP or HPO Vocabulary

The comparison with a human annotator on the PASC symptom section of long hauler sections. Among a total of 1500 annotations:
  ○ 1,022 (179 unique normalized text strings) were annotated by both.
  ○ 414 (251 unique normalized text strings) annotated by NLP but missed by human. Some are human omissions. Some are contextual interpretation.
  ○ 62 (48 unique) annotated by human but missed by NLP. About 8 out of the 48 unique annotations were not in either OMOP or the original PASC concepts. The remaining were caused by lexical variants or multi-span concept mentions (smell was complete lost).
Join the journey for NLP-empowered RECOVER

Hongfang Liu and Emily Plaff
N3C Enclave NLP Data Contribution

[Diagram]

Color Legend (links to scripts below this diagram)
Blue: Use OMOP CREATE TABLE scripts
Green: Use (or adapt) ETL scripts from Mayo
Purple: Use NLP processing code from Mayo
Orange: Use N3C's extract code

https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition/wiki/NLP-Submission-Process
Sites who are interested in deploying the infrastructure, reach out to Katelyn Cordie

(cordie.katelyn@mayo.edu)

Sites who are interested in contributing to the NLP releases, reach out to Sijia Liu

(liu.sijia@mayo.edu)

N3C domain teams who have new NLP needs, reach out to Rafael Fuentes

(rafael.fuentes@nih.gov)

All other questions, reach out to Hongfang Liu

(liu.hongfang@mayo.edu)
• Partnering with TriNetX
• NLP Knowledge Authoring and Concept Mapping Service empowered by Terminology Service and Large Language Models
• Incorporating Open Source NLP Algorithms available for:
  ○ Family History
  ○ Genetic Information
  ○ Context Classifier
  ○ Precision Oncology
  ○ SDoH Variables
Ongoing Community Collaboration

**OHNLP Collaboration**

- Open source tools and resources

**Refinement and Evaluation**

- Continuous improvement

**Context Classification**

- Documentation heterogeneity

**Data Quality and Machine Learning**

- Evaluate the contribution of NLP to data quality and research

#jointhejourney
Acknowledgement

- National Institute of Health
- The Mayo NLP Program
- The N3C NLP Community
- The OHNLP Community
- The CD2H Community
- The CTSA/iEC Community
- The OHDSI Community

Join the Journey in Translating Technology Innovation to Empower Clinical Research towards Better Health for Everyone.
Join the journey

Onboarding to N3C: [bit.ly/cd2h-onboarding-form](bit.ly/cd2h-onboarding-form)

Joining Workstreams:

- **N3C Data Ingestion & Harmonization Workstream**
  - Slack Channel Harmonization
  - Google Group Harmonization

- **N3C Phenotype & Data Acquisition Workstream**
  - Slack Channel Phenotype
  - Google Group Phenotype

- **N3C Collaborative Analytics Workstream**
  - Slack Channel Analytics
  - Google Group Analytics

- **N3C Data Partnership & Governance Workstream**
  - Slack Channel Governance
  - Google Group Governance

- **N3C Synthetic Clinical Data Workstream**
  - Slack Channel Synthetic
  - Google Group Synthetic

Additional Information:
Onboarding N3C, Slack, Google | Finding and Joining a Google Group