Deployment of an OMOP CDM-compatible NLP system for Rapid Development and Dissemination of a Long-COVID Sign/Symptom Extraction NLP task

Andrew Wen*, Liwei Wang*, Huan He, Sunyang Fu, Sijia Liu, Hongfang Liu†, on behalf of the RECOVER team

* These authors contributed equally

† Author to whom correspondence should be addressed

Background

The wealth of longitudinal clinical data contained within the Electronic Health Record (EHR) has given rise to tremendous opportunity for both clinical research and practice. A key component to enabling leveraging this data for research in a cross-institutionally portable manner is the Observational Health Data Sciences and Informatics collaborative, in particular with its Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) [1]. This standard model for representing clinical data has successfully been leveraged to enable rapid execution of many clinical phenotyping tasks across multiple healthcare institutions with heterogeneous EHR systems.

While tremendous effort has been made in ingesting structured portions of EHR data into the OMOP CDM, usage of unstructured data, particularly clinical narratives, in the OMOP CDM is less well defined. Despite the existence of a plethora of clinical natural language processing pipelines, data ingestion and output to these applications vary greatly. While the OMOP CDM offers some standardization to this process as part of the NOTE and NOTE_NLP tables [2], integration with existing NLP solutions is limited. As such, neither the ease of adoption of such an integrated NLP solution amongst OMOP sites nor the ability of the existing CDM tables to meet information needs for practical usage are well understood. In order to evaluate the latter problem, the former must first be addressed such that OMOP-compliant NLP data can be processed across a wide variety of sites.

As part of prior work, we developed an NLP framework integrated with the OMOP CDM, the Open Health Natural Language Processing Toolkit (OHNLP toolkit) [3]. In this work, we will report on the usage of this toolkit to rapidly develop and prototype an NLP system using this toolkit to execute a clinical phenotyping task to meet a real world information need in the form of extracting signs and symptoms related to long COVID.

Methods

Due to PASC's international recognition, many useful literature resources have been developed. A recent literature review[4] identified 303 articles published before April 29, 2021, curated 59 relevant manuscripts that described clinical manifestations in 81 cohorts of individuals three weeks or more following acute COVID-19, and mapped 287 unique clinical findings to Human Phenotype Ontology (HPO) terms. Another study[5] consolidated 355 long COVID symptoms from 1,520 UMLS concepts of 16,466 synonyms, which were identified from 328,879 clinical notes of 26,117 COVID-19 positive patients from their post-acute infection period (day 51-110 from first positive COVID-19 test).

These resources were used as a base concept list, which was then cross-referenced with the Unified Medical Language System (UMLS) version 2021AB to obtain an algorithmically-derived base dictionary.

Lexical variant generation was then ran on these terms to obtain normalized forms and used as a base NLP system. The coverage of this NLP system was then compared against an annotated 98 document clinical narrative set from the Post COVID Care Clinic at the Mayo Clinic, and NLP system finetuning was done using this narrative set to obtain the NLP system used for distribution across the RECOVER team.

This NLP system was then distributed alongside the OHNLP toolkit to 10 RECOVER sites for deployment and federated evaluation. A subset of RECOVER sites (UKen, UMN, Stony Brook, UMichigan, and Columbia) returned evaluation results to the Mayo Clinic to determine dictionary coverage.

Results

In Table 1, we present the results from manual annotation after adjudication as well as the dictionary coverage for these annotations. In Table 2, we present an analysis of the annotations not covered by the NLP algorithm

Table 1. Statistics of dictionary coverage

Results	**UKen**	**UMN**	**Stony Brook**	**Michigan**	**Columbia**
No. of annotation	126	23	171	118	73
No. in dictionary	77	20	138	84	65
Coverage ratio	61.1%	87.0%	80.7%	71.2%	89%

Table 2. Statistics of annotations not covered by dictionary

Results	**UKen**	**UMN**	**Stony Brook**	**Michigan**	**Columbia**
No. of annotations not covered by NLP	49	3	33	34	8
Add to concept	55.1%, 27	66.7%, 2	27.3%, 9	11.8%, 4	12.5%, 1
Add to variant	40.8%, 20	33.3%, 1	51.5%, 17	76.5%, 26	50.0%, 4
Annotation error	4.1%, 2	0	21.2%, 7	11.8%, 4	37.5%, 3

Conclusion

The usage of a common data model for data ingest/output in the NLP pipeline greatly facilitated deployment of the developed NLP system across a wide variety of clinical healthcare sites. Additionally, our results fundamentally demonstrate the advantage of this high degree of portability when discussing generalizable NLP solutions: while the initial Mayo-developed NLP system performs decently across the RECOVER sites that returned evaluation results, not all information needs are met in terms of dictionary coverage. Refinement is thus vastly simplified as gaps in data coverage is assessed across a variety of differing healthcare institutions, thus granting greater confidence in the wide-range generalizability of the final solution.

While the OHNLP toolkit is only one example of how NLP systems can prospectively integrate with the OMOP CDM, we believe that the rapid deployment and ease of refinement demonstrates the importance of such solutions. Additionally, as many sites are now providing this data centrally back to the N3C collaborative for long COVID research use cases, we can begin examining the problem of whether the existing CDM tables NOTE and NOTE_NLP meet information needs for practical usage.

References

1. Observational Health Data Sciences and Informatics [Internet]. OHDSI. [cited 2022Jun24]. Available from: https://www.ohdsi.org/
2. OMOP CDM v5.3. [cited 2022Jun24]. Available from: https://ohdsi.github.io/CommonDataModel/cdm53.html#NOTE
3. Liu S, Wen A, Wang L, He H, Fu S, Miller R, et al. An Open Natural Language Processing Development Framework for EHR-based Clinical Research: A case demonstration using the National COVID Cohort Collaborative (N3C). arXiv:2110.10780
4. Deer RR, Rock MA, Vasilevsky N, Carmody L, Rando H, Anzalone AJ, et al. Characterizing long covid: Deep phenotype of a complex condition. eBioMedicine. 2021;74:103722.
5. Wang L, Foer D, MacPhaul E, Lo Y-C, Bates DW, Zhou L. Pasclex: A comprehensive post-acute sequelae of Covid-19 (PASC) symptom lexicon derived from electronic health record clinical notes. Journal of Biomedical Informatics. 2022;125:103951.