An Open Natural Language Processing Development Framework for EHR-based Clinical Research: A case demonstration using the National COVID Cohort Collaborative (N3C)

Sijia Liu1,*, Andrew Wen1,*, Liwei Wang1,*, Huan He1,*, Sunyang Fu1, Robert Miller2, Andrew Williams2, Daniel Harris3, Ramakanth Kavuluru3, Mei Liu4, Noor Abu-el-rub4, Rui Zhang5, Masoud Rouhizadeh6, John D. Osborne7, Yongqun He8, Umit Topaloglu9, Stephanie S Hong10, Joel H Saltz11, Thomas Schaffer12, Emily Pfaff13, Christopher G. Chute10, Tim Duong14, Melissa A. Haendel15, Rafael Fuentes16, Peter Szolovits17, Hua Xu18, Hongfang Liu1, National COVID Cohort Collaborative (N3C) Natural Language Processing (NLP) Subgroup, National COVID Cohort Collaborative (N3C)

1. Department of Artificial Intelligence and Informatics, Mayo Clinic.
2. Tufts Clinical and Translational Science Institute, Tufts Medical Center.
3. Department of Internal Medicine, University of Kentucky.
4. Department of Internal Medicine, University of Kansas Medical Center.
5. Department of Pharmaceutical Care & Health Systems, University of Minnesota at Twin Cities.
6. Department of Pharmaceutical Outcomes & Policy, University of Florida
7. Department of Computer Science, University of Alabama at Birmingham.
8. University of Michigan Medical School.
9. Wake Forest Baptist Medical Center.
10. Department of Medicine, Johns Hopkins University.
11. Department of Biomedical Informatics, Stony Brook University.
12. Sage Bionetwork.
Abstract

While we pay attention to the latest advances in clinical natural language processing (NLP), we can notice some resistance in the clinical and translational research community to adopt NLP models due to limited transparency, interpretability, and usability. In this study, we proposed an open natural language processing development framework. We evaluated it through the implementation of NLP algorithms for the National COVID Cohort Collaborative (N3C). Based on the interests in information extraction from COVID-19 related clinical notes, our work includes 1) an open data annotation process using COVID-19 signs and symptoms as the use case, 2) a community-driven ruleset composing platform, and 3) a synthetic text data generation workflow to generate texts for information extraction tasks without involving human subjects. The corpora were derived from texts from three different institutions (Mayo Clinic, University of Kentucky, University of Minnesota). The gold standard annotations were tested with a single institution's (Mayo) ruleset. This resulted in performances of 0.876, 0.706, and 0.694 in F-scores for Mayo, Minnesota, and Kentucky test datasets, respectively. The study as a consortium effort of the N3C NLP subgroup demonstrates the feasibility of creating a federated NLP algorithm development
and benchmarking platform to enhance multi-institution clinical NLP study and adoption. Although we use COVID-19 as a use case in this effort, our framework is general enough to be applied to other domains of interest in clinical NLP.

Introduction

Over the past decade, Electronic Health Record (EHR) systems have been increasingly implemented at US hospitals and clinics. Large amounts of detailed longitudinal patient information, including lab tests, medications, disease status, and treatment outcomes, have been accumulated and are available electronically. These large clinical databases are valuable data sources for clinical and translational research. Major initiatives have been established to exploit this crucial resource, including the Clinical and Translational Science Awards (CTSA) Clinical Data to Health (CD2H) program [1], the Electronic Medical Records and Genomics (eMERGE) Network [2], the Patient-Centered Outcomes Research Institute’s (PCORI) Clinical Research Networks (CRNs), the NIH All of Us Research Program, and the Observational Health Data Science and Informatics (OHDSI) Consortia. Such EHR-based networks have greatly accelerated biomedical research. Tremendous efforts and resources have been devoted to developing infrastructures for secondary use of EHRs data and have resulted in significant discoveries in clinical and translational research.

Prior to the 1980s, the development of NLP systems involved handcrafting large sets of language processing knowledge bases (i.e., symbolic NLP). Several early attempts relied on corpus analysis conducted manually by experts, e.g., Linguistic String Project (LSP) at NYU and Medical Language Extraction and Encoding (MedLEE) system [3]. However, because of the high variability of clinical subdomains and information sources, symbolic NLP systems were very expensive to construct and sometimes criticized for lack of portability. Since 2000, modern approaches have been grounded in statistical inference (i.e., statistical NLP) through analysis of
large corpora of real-world annotated examples. Since 2017, the use of large pre-trained neural language models has become the model of choices for NLP tasks.

While we pay attention to the latest advances in clinical NLP, we also notice some resistance in the clinical and translational research community to adopt NLP techniques which can be summarized in the following:

**Interoperability/Usability/Transparency.** A general clinical NLP system often consists of a concept mention detection component that recognizes mentions of important clinical concepts such as diseases and drugs, and a concept encoding component that maps clinical text to concepts in standard terminologies, thus allowing seamless integration with other computerized applications that rely on terminology standards. Although substantial efforts have been devoted to clinical information extraction, the performance of existing clinical NLP systems for concept encoding is not satisfactory as concept encoding itself is a challenging task even for domain experts due to the lack of common consensus. The existing gold standard annotations are subject to semantic rigidity. Meanwhile, there is significant variability across various EHR-based studies regarding concept definition. The data heterogeneity caused by the variety associated with EHR implementation brings additional challenges in implementation. EHR system variations and associated healthcare systems’ modes of operation also typically lead to idiosyncratic ways of clinical documentation. All these diminish the reproducibility of NLP-empowered clinical studies where NLP primarily serves as a method for data abstraction. A transparent process is needed to ensure reproducibility in NLP-empowered clinical studies.

**Explainability/Interpretability/Implementability.** Generally, NLP methodologies can be divided into expert-based or machine learning-based approaches (i.e., statistical NLP). Expert-based methods are easily tailored to a dataset, but statistical NLP is relatively more portable. While statistical NLP shows high performance for many tasks, to be viable in a real-world clinical setting, it requires an annotated corpus to be created after addressing any errors found need to be determinable and implementable without a significant investment of effort. In the clinical domain,
creating such annotated corpora requires significant domain expertise and time to produce. This demand for large corpus annotations may be partially mitigated by the adoption of transfer learning or distant supervision with pre-trained language models. However, interpretability and explainability become nontrivial barriers to implementation and uptake of NLP applications [4]. Specifically, interpretability involves coming up with clinician-friendly rationales for NLP system outputs. Explainability deals with ways to reliably link the innards of an NLP model to outputs in an instance-specific manner.

**Usability.** Another common hurdle in the adoption of NLP solutions is the usability and adaptability, particularly in organizations without NLP expertise. Meanwhile, tolerance for error is very low in the clinical domain. When considering error resolution in clinical NLP, an issue becomes evident: fixing rules is easy, but, because many statistical NLP models are inherently neither interpretable nor explainable to humans, errors are difficult to debug without an understanding of the statistical model. Interpretability is the bigger hurdle here in terms of real-world usage of NLP systems.

Built upon our previous work, in this study, we proposed an open natural language processing development framework and evaluate it through the implementation of NLP algorithms for the National COVID Cohort Collaborative (N3C). The issues identified earlier are addressed by the following components: 1) A NLP system with clinical common data models; 2) A crowdsourcing development tool based on web-based graphical user interface (GUI); 3) Multi-site participation on evaluation and corpus development. The proposed framework can incorporate different NLP engines as long as the input and output are governed and standardized by the same specification. In this study, we used a rule-based approach which is based on regular expressions, dictionary lookup, and the ConText algorithm to address with the use case of symptom extraction for COVID-19 [5]. The reason we opt for this solution is its simplicity, transparency, and interpretability as the outcomes are fully deterministic based on the definition of the rules. When a baseline rule and dictionary is made available to the public, it can be updated by different users from different sites.
Results

Crowdsourcing platform

An open-source collaborative platform for developing NLP rulesets has been developed and hosted at https://ohnlp4covid-dev.n3c.ncats.io/. The demo homepage (Figure 2(a)) is to demonstrate the N3C NLP engine outputs on annotating clinical texts using the aforementioned baseline rulesets and dictionary. The annotations are from components of Sign/Symptom extractor, temporal information extractor and dictionary lookup extractor. To further customize each model, the users can visit “Rule Editor” (https://ohnlp4covid-dev.n3c.ncats.io/ie_editor) and the “Dictionary Builder” (https://ohnlp4covid-dev.n3c.ncats.io/dict_builder) page (Figure 2(b)). Figure 2(c) provides an example ruleset named “covid19” which includes all signs and symptoms with the key words for further editing and customization. The rulesets can be tested in real time by clicking the “Upload and test” button, where the rulesets will be uploaded, and the NLP engine will be generated for testing and debugging.

Distributed text data collection

The N3C deidentification and synthetic text generation workflow is illustrated in Figure 1. Clinical notes from patients with positive COVID-19 test results from 2 other institutions (the University of Kentucky, the University of Minnesota at Twin Cities) were initially collected. Notes that were not office visit notes (e.g., nurse calls, etc.), notes that had fewer than 1000 characters, and notes that were authored more than 14 days prior to the date of the patient’s earliest positive COVID-19 test result were further filtered out. A total of 220 clinical notes from these sites that met these criteria were randomly selected for the further de-identification process.
Prior to using the de-identification program developed by the Medical College of Wisconsin, it was necessary to make modifications to the code so the program would work with the SQL database available in the secure data environment. The date offset was set to a random integer between -15 and +15 days. The university name and the names of some of the major health systems in the area were added to the blacklist file. Afterwards, the deidentified notes were manually reviewed to identify and remove any additional PHI that the program may have missed. Once the manual review process was completed, an ID was generated and assigned to each note so that the ID generated by the EHR system could be removed from the final dataset.

Centralized annotation of COVID-19 signs and symptoms

We collected 20 signs and symptoms of COVID-19 as a basic COVID-19 concept set according to the recommendations from the CDC and Mayo Clinic. Five out of the 20 concepts are emergency warning signs including dyspnea, chest pain, delirium, hypersomnia and cyanosis. We then gathered formal definitions of each clinical concept from the Coronavirus Infectious Disease Ontology (CIDO) [7]. Based on the Open Biological and Biomedical Ontology (OBO) Foundry library, CIDO concepts were imported from 45 ontologies, and it uses Human Phenotype Ontology (HPO) [8] for phenotypes. Some representative phenotypes shown in COVID-19 have been imported to the CIDO. However, if the chosen COVID-19 clinical concepts were not collected by the CIDO, we re-pulled them from the HPO to the CIDO. We also gathered cross-reference concept codes from the CIDO including UMLS [9], SNOMED-CT [10], MeSH [11], HPO, MeDDRA [12].

We selected all available clinical notes from both inpatients and outpatients in the two-week window preceding the order date of the first positive COVID-19 result as the annotation cohort. After the text data was collected from participating sites, the same annotation process was completed by the annotator team from Mayo Clinic to generate the gold standard annotations on COVID-19 signs and symptoms. There are 313 clinical notes from Mayo Clinic, 20 notes from the
University of Kentucky (UKen) and 36 notes from University of Minnesota (UMN). Automatic PHI de-identification pre-process was performed before manual annotation. To detect any PHI missed by the automatic de-identification process, additional manual annotation of PHI was done in parallel with the sign and symptom annotation. The annotated PHI information was then post-processed as a part of the de-identification process. Inter annotator agreement was calculated after annotation (0.686 F1-score) and corresponding discrepancies were resolved to generate a final gold standard dataset.

**NLP algorithm development**

Using the annotated corpus, we developed a single-site NLP algorithm and a multi-site NLP algorithm using a regular expression-based matching method, which has been widely adopted for information extraction in clinical settings [13, 14]. Specifically, for the Mayo data, we randomly chose 101 notes out of the 313 annotated notes as development set, 105 notes as validation set, and the remaining 107 notes were used as the testing set. For the UKen data, 10 notes were used as training and 10 as testing. For the UMN data, 18 was used as training and 18 as testing. The development set and validation set from Mayo were used for training and refining both single-site algorithm and multi-site algorithm, that was tested on all data from UKen and UMN for the single-site algorithm, and further improved by using training sets from UKen and UMN and then tested on testing sets from all sites.

**Evaluation**

We evaluated the algorithms developed from single-site data and multi-site data. The performance is measured in precision, recall, and F1-score of the annotated concepts, without and with certainty. A span is defined as the start position and the end position for the concept. For the span detection, when there are overlaps between the gold standard mention and detected concept and the concept type (i.e., the sign/symptom) is the same, it is considered a true positive (TP). If a concept exists
in the gold standard annotation but not detected by the algorithm, or the concept type is not matched, is considered as a false negative (FN). If a concept is detected by the algorithm but does not exist in the gold standard annotation, the concept is considered as a false positive (FP). The precision, recall and F1-score are then calculated as follows:

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

For the concept extraction with span and certainty evaluation, the distinction is that it requires the span, the certainty needs to be matched in addition to the concept type to be considered as a true positive. Table 1 shows the performance of the single-site NLP algorithm and Table 2 shows the performance of the multi-site NLP algorithm.

Sampled error analysis found that major discrepancies in certainty between NLP algorithm and the gold standard were due to patient education, tokenization error due to de-identification process, semi-structured template, the wrong gold standard, and other note sections such as medication.

**Discussion**

**Continuous development**

Since the NLP models to be shared in rule-based systems can be decomposed to eliminate the possibility of leaking PHI, both the synthetic corpora and models can be shared without considering these language resources involves human subjects. Therefore, it can mitigate some privacy constraints in multi-site unstructured data sharing.
In the proposed workflow, each site annotates some results of their NLP systems using the community defined concepts by adopting annotation and data curation guidelines. The curated datasets can also be used in a cloud computing setting, where the distributed evaluation on each dataset from each participating site is handled by a centralized controller. NLP Sandbox is an example of such evaluation framework, which uses docker containers to encapsule algorithm implementations. By adopting this process, there is no need for acquiring data for federated evaluation, since the evaluation only happens within each institution, and no data has been transferred out of the data providing site. The performances regarding the precision, recall and F1-scores defined in various experimental settings can be obtained in near real-time and thus can be used into the continuous development workflows.

N3C domain task teams

The N3C NLP platform also enables augmenting centralized structured data in N3C Enclave with the clinical concepts from clinical narratives [15]. Some ongoing initiatives include composing rule-based algorithms for Long COVID signs and symptoms [16] and outcomes of COVID-19 in patients with cancer [17]. The framework we propose can be leveraged in a straightforward manner to facilitate this augmentation of the N3C dataset. As it is well known that social determinants of health are often discussed in free text, this process can improve the comprehensiveness of patient-specific data made available through the N3C.

Interoperability

Efforts from the clinical standard communities like OMOP and FHIR have been made to enhance the interoperability of text data. One direction is to standardize clinical concepts from narrative clinical texts into terms defined by standard vocabularies such as SNOMED CT or ICD-9-CM [18, 19]. As the methods are developed via publicly available models and language resources, the systems are interoperable thus portable among different institutions.
Methods

Framework Description

The system consists of data ingestion layer, processing layer and data persistence layer. The architecture of the proposed framework is illustrated in Figure 3. In the framework that does not support OMOP CDM, the data ingestion layer works as the data collector to read texts from various sources such as relational databases and file systems to load them into the NOTE table of OMOP CDM. The framework keeps both the document metadata and the contents and makes both searchable via loading them into ElasticSearch. The processing layer serves as the NLP engine where the information extraction from raw texts happens given its heuristic rules created by various NLP engines like MedTagger [20]. After the term modifiers added by contextual rules from ConText Algorithm [5] around the extracted condition mentions, these conditions will compose clinical events with temporal information. The clinical events will be stored on OMOP CDM NOTE_NLP table as the events extracted from NLP systems before distributed to other tables of OMOP CDM (e.g., CONDITION_OCCURRENCE, MEASUREMENT, DRUG_EXPOSURE).

The system is distributed as an open-source software under Apache 2.0 license via Github. An extraction script is also developed to prepare the NOTE_NLP table for N3C Enclave submission by removing data elements with risks of protected health information (PHI) for the distribution to N3C Enclave. As a result, the elements retained only contain deidentified document and patient metadata and codified OMOP concepts from clinical vocabularies. The elements with original texts such as text snippets and concept mentions are truncated before submission. The system has been deployed in 4 institutions and these sites are routinely contributing NLP data to N3C Enclave.
Incorporating standards and interoperability

Clinical concept standardization is a critical component in achieving the semantic interoperability for the secondary use of EHRs, e.g., unstructured data through NLP and structured data. An ontology is a formal, explicit description of concepts, their properties, relationships between the concepts, and the allowed values that they may take. In this study, we developed basic COVID-19 clinical concepts and standardized these concepts through our collaboration with the CIDO.

A common barrier to the widespread adoption of NLP in clinical research is the need to transform input and outputs to conform to part of an overall pipeline. While seemingly straightforward, such a task is difficult without prior significant investment in associated infrastructure and dedicated software development. It is therefore desirable to leverage existing infrastructure where possible and incorporate such an effort into the distributed NLP pipeline to reduce technical burden on the end user.

There is, however, significant variation in terms of available infrastructure and data availability amongst different institutions. Creating a solution that is immediately suitable for all these environments out of the box would be immensely challenging. For that reason, we sought to leverage existing data modeling efforts that are likely to be already adopted by academic medical institutions to standardize the data ingestion and output process. In our implementation, we chose the Observational Health Data Sciences and Informatics’ Observational Medical Outcomes Partnership common data model (OHDSI/OMOP CDM) to handle input of clinical narratives via the NOTE table and output via the NOTE_NLP table. This brings the advantage that input/output is now standardized: so long as institutions have already transformed their clinical data into the OMOP CDM, and/or their downstream NLP-reliant applications read from the OMOP CDM database, no additional technical development burden is needed.

It is important to note that standardization as a default only serves to simplify adoption for those who already have a solution complying with the standard and cannot be a comprehensive solution.
A purely OMOP CDM reliant solution is not ideal, as not all institutions will have their own OMOP CDM instance and standing up such an instance to just use a pipeline may produce undue burden. For that reason, input/output in our infrastructure is modularized, and can be substituted at will: the default OMOP CDM I/O utilizes a variant of SQL-based data extractors/writers, and the specific query and connection strings used can be substituted via plaintext configuration changes. Additionally, SQL-based I/O is not the only supported setting, a variety of other data sources including Elasticsearch, google cloud storage, amazon s3, and plaintext are included as well as configuration-swappable options.

Crowdsourcing approach

To promote collaboration and the sharing efforts of the participation in the algorithm development, we build a crowdsourcing platform for the domain experts to upload, customize, and examine their NLP algorithms in an interactive web application. Users could create keyword-based and rule-based algorithms and test the performance in the online environment instantly. The crowdsourcing platform consists of three modules based on our ETL and NLP system to support expert collaboration, including dictionary builder, regular expression rule set editor, and detection result visualization.

The dictionary builder could extend the keyword collection used by the algorithm. Users could customize particular terms from the ontology database such as CIDO [7] and MONDO [21]. The regular expression rule set editor provides an integrated interface to help users customize their own regular expression rule set based on existing solutions to optimize the specific performance, such as new symptoms, treatments, or outcomes. The detection result visualization is designed based on Brat annotation tool [22] to check the results generated by different methods.
Deidentification and Synthetic dataset curation

To protect the patient's privacy according to HIPAA Privacy Rule [23], a set of identifiers and protected health information needs to be removed. This does not require removal of general health conditions in clinical texts. However, to further prevent tracing of patient records and to generate synthetic datasets which may not necessarily be considered as data related to human subjects. We used notes deidentification software to identify and replace PHI with their meaningful substitutes. For example, a list of person and hospital names are curated to replace the real patient and hospital names in the clinical texts, and each digit in phone numbers is replaced by a random digit.
Acknowledgment

This research was possible because of the patients whose information is included within the data and the organizations and scientists who have contributed to the on-going development of this community resource https://doi.org/10.1093/jamia/ocaa196. The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v1.2-2020-08-25b and supported by NCATS U24 TR002306. This task was made possible by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number U01TR02062 and the Bill & Melinda Gates Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

Data curation: SL, AWen, LW, HH, RM, Awilliams, DH, RK, ML, NA, MR, RZ, JDO, JHS
Data integration: AWen, LW, HH, RM, DH, RK, ML, NA, RZ, TS, YH, EP, SSH, CGC, JHS
Data analysis: SL, AWen, LW, RM, RK, NA, RZ, MR, TS
Software development: SL, AWen, HH, MR, TS
Data quality assurance: SL, RM, DH, RK, LM, NA, RZ, TS, JDO, HY, EP, TD, PS, HX
Draft the manuscript: SL, AWen, LW, HH, RM, RZ, HL
Critical revision of the manuscript for important intellectual content: SF, Awilliams, RK, ML, NA, YH
Project evaluation: LW, SF, RM, Awilliams, DH, RK, ML, NA, RZ, MR, TS
Project management: SL, LW, RZ, TS, EP, JHS, RF, HL
Regulatory oversight / admin: EP, CGC, HL

Database / Information systems admin: RM, DH, NA, RZ, TD

Biological subject matter expertise: LW, YH, UT, MAH

Funding acquisition: MAH, CGC, PS, HX, HL
Reference

1. *Welcome | Center for Data to Health*. Available from: https://cd2h.org/mission.

2. Gottesman, O., et al., *The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future*. Genetics in Medicine, 2013. 15(10): p. 761-771.

3. Friedman, C., et al., *Automated encoding of clinical documents based on natural language processing*. Journal of the American Medical Informatics Association : JAMIA, 2004. 11(5): p. 392-402.

4. Barredo Arrieta, A., et al., *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*. Information Fusion, 2020. 58: p. 82-115.

5. Harkema, H., et al., *ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports*. Journal of biomedical informatics, 2009. 42(5): p. 839-851.

6. *MCW_BMI / notes-deidentification*. Available from: https://bitbucket.org/MCW_BMI/notes-deidentification/src/master/.

7. He, Y., et al., *CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis*. Scientific Data, 2020. 7(1): p. 181.

8. Köhler, S., et al., *The Human Phenotype Ontology in 2021*. Nucleic Acids Res, 2021. 49(D1): p. D1207-d1217.

9. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology*. Nucleic Acids Research, 2004. 32(90001): p. 267D-270.

10. *SNOMED-CT*.

11. Medicine, N.L.o. *Medical Subject Headings*. Available from: https://www.nlm.nih.gov/mesh/meshhome.html.

12. Brown, E.G., L. Wood, and S. Wood, *The medical dictionary for regulatory activities (MedDRA)*. Drug Saf, 1999. 20(2): p. 109-17.

13. Wang, Y., et al., *Clinical information extraction applications: A literature review*. Journal of Biomedical Informatics, 2018. 77.
14. Fu, S., et al., Clinical concept extraction: A methodology review. Journal of Biomedical Informatics, 2020. 109: p. 103526.

15. Haendel, M.A., et al., The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association, 2021. 28(3): p. 427-443.

16. Rando, H.M., et al., Challenges in defining Long COVID: Striking differences across literature, Electronic Health Records, and patient-reported information. medRxiv, 2021: p. 2021.03.20.21253896.

17. Sharafeldin, N., et al., Outcomes of COVID-19 in Patients With Cancer: Report From the National COVID Cohort Collaborative (N3C). Journal of Clinical Oncology, 2021: p. JCO.21.01074.

18. Peterson, K.J., G. Jiang, and H. Liu, A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. Journal of Biomedical Informatics, 2020. 110: p. 103541.

19. Yuan, C., et al., Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc, 2019. 26(4): p. 294-305.

20. Github - OHNLP/MedTagger. Available from: https://github.com/OHNLP/MedTagger.

21. Mungall, C.J., et al., The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Research, 2017. 45(D1): p. D712-D722.

22. Stenetorp, P., et al. brat: a Web-based Tool for NLP-Assisted Text Annotation. 2012. Avignon, France: Association for Computational Linguistics.

23. Services, U.S.D.o.H.H., The HIPAA Privacy Rule.
Figures

Figure 1. N3C deidentification and synthetic text generation workflow
Figure 2. Screenshots of the Web GUI

(a)
DYSPEA
- Gastroesophageal reflux disease
- Excessive stomach acid
- Nausea, vomiting, and bloating
- Difficult or painful swallowing

NASAL OBSTRUCTION
- Blocked nose
- Congested nose
- Sinus pressure
- Difficulty breathing
- Snoring

LOSS OF APETITE
- Loss of appetite
- Appetite loss
- Food aversion

LYMPHOPENIA
- Decreased lymphocytes
- Low lymphocyte count
- Low lymphocyte number

ELEVATED LDH
- Elevated lactate dehydrogenase
- Increased LDH
- Increased lactate dehydrogenase activity

ELEVATED PT TIME
- Prolonged prothrombin time
- Abnormal prothrombin time
- Abnormal PT
The tree is built on 6818 classes:

- All
  - sequence feature
  - entity
  - continuous
  - document
  - process
    - host response to coronavirus infection
    - disorder prevention
    - infectious disease epidemic
    - vaccine-induced host response
    - exposure event or process
    - physiological effects (PE)
    - life cycle
    - host exposure to infectious agent
    - coronal process to host
    - life-death temporal boundary
    - cellular or molecular interactions (MMIC)
    - immunization against infectious agent
    - immune response
    - host coronavirus interaction
    - immunization
    - life-cycle stage
    - clinical kinetics (PK)
    - ...
Figure 3. Diagram of N3C NLP Solution
# Tables

Table 1. Performance of the single-site NLP algorithm (Mayo: Mayo Clinic, UKen: University of Kentucky, UMN: University of Minnesota)

| Dataset | Span   | Span+Certainty | # entities |
|---------|--------|----------------|------------|
|         | Precision | Recall | F1 | Precision | Recall | F1 |          |
| Mayo    | 0.882   | 0.869 | 0.876 | 0.789 | 0.639 | 0.706 | 146      |
| UKen    | 0.698   | 0.714 | 0.706 | 0.664 | 0.643 | 0.653 | 46       |
| UMN     | 0.658   | 0.735 | 0.694 | 0.534 | 0.438 | 0.481 | 150      |

Table 1. Performance of the multi-site NLP algorithm (Mayo: Mayo Clinic, UKKentucky: University of Kentucky, UMN: University of Minnesota)

| Dataset | Span   | Span+Certainty | # entities |
|---------|--------|----------------|------------|
|         | Precision | Recall | F1 | Precision | Recall | F1 |          |
| Mayo    | 0.863   | 0.908 | 0.884 | 0.824 | 0.681 | 0.746 | 146      |
| UKen    | 0.696   | 0.859 | 0.769 | 0.662 | 0.734 | 0.696 | 46       |
| UMN     | 0.718   | 0.918 | 0.806 | 0.562 | 0.456 | 0.504 | 150      |