Summer 8-2021

Standardizing New Diagnostic Tests to Facilitate Rapid Responses to The Covid-19 Pandemic

Xiao Dong
University of Texas Health Science Center at Houston, Xiao.Dong@uth.tmc.edu

Standardizing New Diagnostic Tests to Facilitate Rapid Responses to The Covid-19 Pandemic

By

Xiao Dong, MD, MS

APPROVED:

DocuSigned by:
Hua Xu
53CDB41ACE9A4FD...
Hua Xu, PhD, Chair

DocuSigned by:
Amy Franklin
263A215A76DA461...
Amy Franklin, PhD

DocuSigned by:
YangDong
E828DA81F0BF458...
Yang Gong, PhD

DocuSigned by:
Cui Tao
89F39B9A1BE943C...
Cui Tao, PhD

Date Approved: 08/10/2021

Standardizing New Diagnostic Tests to Facilitate Rapid Response to The Covid-19 Pandemic

A
Dissertation

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

By

Xiao Dong, M.D., M.S.

University of Texas Health Science Center at Houston

2021

Dissertation Committee:

Hua Xu, PhD[1], Advisor
Amy Franklin, PhD[1]
Yang Gong, PhD[1]
Cui Tao, PhD[1]

---

[1]The School of Biomedical Informatics

Dedication

Dedicated to Huakang and Steven Tu.

Acknowledgements

I am truly thankful to my primary advisor, Dr. Hua Xu, for his dedicated and brilliant mentorship since the beginning of my PhD study. Without his guide, I could not smoothly switch from a background of medical practicing to the biomedical informatics field. I would also like to thank my committee members: Dr. Cui Tao, Dr. Yang Gong, and Dr. Amy Franklin, who have been extremely supportive during my dissertation topic selection and project completion. They also provided critical help to address the challenges in the dissertation research.

I am also appreciative of the mentors and staff involved in the pre-doctoral fellowship in Innovation for Cancer Prevention Research (CPRIT). Specifically, I would like to thank Dr. Patricia Mullen, Dr. Roberta Ness, and Dr. Sahiti Myneni for their mentoring.

I would like to acknowledge my colleagues in Dr. Xu and Dr. Tao's lab: Dr. Yaoyun Zhang, Dr. Qiang Wei, Dr. Jingcheng Du, Jingqi Wang, Jianfu Li, Yujia Zhou, Xu Zuo, for their help in my doctoral studies. I would also like to give thanks to my friends and collaborators at UTHealth SBMI for all their assistance.

In addition, I would like to thank CPRIT RP160015 for their support in this study.

Abstract

In order to enhance the data interoperability, an expeditious and accurate standardization solution is highly desirable for naming rapidly emerging novel lab tests, and thus diminishes confusion in early responses to pandemic outbreaks. This is a preliminary study to explore the roles and implementation of medical informatics technology, especially natural language processing and ontology methods, in standardizing information about emerging lab tests during a pandemic, thereby facilitating rapid responses to the pandemic. The ultimate goal of this study is to develop an informatics framework for rapid standardization of lab testing names during a pandemic to better prepare for future public health threats. We first constructed an information model for lab tests approved during the COVID-19 pandemic and built a named entity recognition tool that can automatically extract lab test information specified in the information model from the Emergency Use Authorization (EUA) documents of the U.S. Food and Drug Administration (FDA), thus creating a catalog of approved lab tests with detailed information. To facilitate the standardization of lab testing data in electronic health records, we further developed the COVID-19 TestNorm, a tool that normalizes the names of various COVID-19 lab testing used by different healthcare facilities into standard Logical Observation Identifiers Names and Codes (LOINC). The overall accuracy of COVID-19 TestNorm on the development set was 98.9%, and on the independent test set

was 97.4%. Lastly, we conducted a clinical study on COVID-19 re-positivity to demonstrate the utility of standardized lab test information in supporting clinical research. We believe that the result of my study indicates great a potential of medical informatics technologies for facilitating rapid responses to both current and future pandemics.

Vita

January 2016 - present ……Ph.D. candidate/student, Biomedical Informatics,
The University of Texas Health Science Center at Houston, Houston, United States

January 2012 – December 2013……Master of Science in Bioinformatics,
Georgia Institute of Technology, Atlanta, United States

September 2007 – July 2010 ……Master of Medicine in Neurology,
Peking University First Hospital, Beijing, China

September 2002 – July 2007……Bachelor of Medicine in Medicine (Equivalent to MD),
Peking University, Beijing, China

Publications

1. **Dong X,** Zhou Y, Shu X, Bernstam E, Stern R, Aronoff DM, Xu H, Lipworth L. Comprehensive characterization of COVID-19 patients with repeatedly positive SARS-CoV-2 tests using a large US electronic health record database. (Accepted by *Microbiology Spectrum*)

2. **Dong X**, Li J, Soysal E, Bian J, DuVall S, Hanchrow E, Liu H, Lynch K, Matheny M, Natarajan K, Ohno-Machado L, Pakhomov S, Reeves R, Sitapati A, Abhyankar S, Cullen T, Deckard J, Jiang X, Murphy R, Xu H. COVID-19 TestNorm - A tool to normalize COVID-19 testing names to LOINC codes. *Journal of American Medical Informatics Association.* 2020 Jun 22;ocaa145. doi: 10.1093/jamia/ocaa145. Online ahead of print.
3. Zuo X, Li J, Zhao B, Zhou Y, **Dong X** , Duke J, Natarajan K, Hripcsak G, Shah N, Banda J, Reeves M R, Xu H. Normalizing Clinical Document Titles to LOINC Document Ontology: an Initial Study. *AMIA 2020 Annual Symposium Proceedings*
4. Chen H, Shi L, Xue M, Wang N, **Dong X**, Cai Y, Chen J, Zhu W, Xu H, Meng Q. Geographic variations in in-hospital mortality and use of percutaneous coronary intervention following acute myocardial infarction in China: a nationwide cross-sectional analysis. *Journal of the American Heart Association*. 2018 Apr 11. doi:10.1161/JAHA.117.00813
5. Miao S, **Dong X (co-first author)**, Zhang X, Jing S, Zhang X, Xu T, Wang L, Du X, Xu H, Liu Y. Detecting Pioglitazone Use and Risk of Cardiovascular Events Using Electronic Health Record Data in a Large Cohort of Chinese Patients with Type 2 Diabetes. *Journal of Diabetes*. 2019 Aug;11(8):684-689. doi: 10.1111/1753-0407.12894. Epub 2019 Feb 5.
6. Tu H, Sun L, **Dong X**, Gong Y, Xu Q, Jing J, Bostick R, Wu X, Yuan Y. A serological biopsy using five stomach-specific circulating biomarkers for gastric cancer risk assessment: a multi-phase study. *Am J Gastroenterol*. 2017 Mar 21. doi: 10.1038/ajg.2017.55.
7. **Dong X**, Zhang Y, and Xu H. Search Datasets in Literature: A Case Study of GWAS. *2017 Joint Summits on Translational Science of AMIA*
8. Xu J, Zhang Y, Wu Y, Wang J, **Dong X**, and Xu H. Citation Sentiment

Analysis in Clinical Trial Papers. *AMIA Annu Symp Proc.* 2015.

9. Tu H, Sun LP, **Dong X**, Gong YH, Xu Q, Long Q, Flanders WD, Smith RA, Bostick RM, Yuan Y. Temporal Changes in Serum Biomarkers and Risk for Progression of Gastric Precancerous Lesions: a Longitudinal Study. *Int J Cancer*. 2014 Jun 4. Int J Cancer. 2015 Jan 15;136(2):425-34. doi: 10.1002/ijc.29005.
10. Tu H, Sun LP, **Dong X**, Gong YH, Xu Q, Jing JJ, Yuan Y. Serum anti-Helicobacter pylori immunoglobulin G titer correlates with grade of histological gastritis, mucosal bacterial density, and levels of serum biomarkers. *Scand J Gastroenterol*. 2014 Mar;49(3):259-66. doi: 10.3109/00365521.2013.869352.
11. Cao K, Lailler N, Zhang Y, Kumar A, Uppal K, Liu Z, Lee E, Wu H, Medrzycki M, Pan C, Ho P, Cooper P, **Dong X**, Bock C, Bouhassira E, Fan Y. High-resolution mapping of h1 linker histone variants in embryonic stem cells. *PLoS Genet*. 2013;9(4)

Field of Study

Health Informatics

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Literature Review

1.1 Introduction

Diagnostic tests play an important role in pandemics such as COVID-19 [1, 2]. They identify recently recognized or emerging microbes that provoke disease outbreaks [3]. Clinicians depend on laboratory tests to diagnose individuals for disease, while public health practitioners monitor threats based on diagnostic test results. During the COVID-19 pandemic, clinical laboratory diagnostics have acted an even more prominent role [1, 2, 4].

The EUA [5] is the mechanism by which the FDA authorizes unapproved medical solutions under public health emergencies. Currently, all the U.S. COVID-19 lab tests are authorized under EUA [5] in order to rapidly support the early response and mitigation. The EUA mechanism ensured rapid laboratory support for the battle with COVID-19 [5]. However, it also posed challenges for standard reporting and efficient analysis of diagnostic results, as many new tests were described in narrative text without a standard representation (i.e., codes in standard terminologies), especially in the early stage of the pandemic [6–8]. Therefore, expeditious and accurate standardization of information about the rapidly emerging and novel lab tests is highly desirable, which will improve the data interoperability and diminish confusion in the early response to pandemic outbreaks.

The significance of medical informatics in combating COVID-19 has been widely acknowledged in various studies [9–14]. However, limited work has discussed the challenges regarding the standardization of names of diagnostic tests, and few informatics tools have been developed to tackle this issue. At present, challenges to standardizing new diagnostic tests in COVID-19 include: (1) there is no standard representation of critical information for each EUA approved diagnostic test and the lack of tools to automatically extract such information; (2) at the beginning of the pandemic, lab tests reported by each healthcare system often did not follow standard terminologies (e.g., LOINC [15] and the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [16]), and how to standardize those tests and their results is crucial, but little work has been done yet; (3) more applications are required to demonstrate the use of normalized diagnostic test results from large populations.

To bridge above identified gaps, we propose an informatics framework based on the natural language processing (NLP) and ontology technologies, to standardize diagnostic test names in EUA documents and electronic health records, thus facilitating clinical studies based on standardized test data. Specific aims of this study include:

Aim 1: Develop an information model and an information extracting tool to automatically identify new diagnostic test information from EUA documents.

We developed a specific information model for new COVID-19 diagnostic tests, which utilizes concepts from LOINC and SNOMED CT together with some original concepts. Based on it, we further developed an information extraction tool to automatically identify relevant entities in FDA EUAs for COVID-19 lab tests. A baseline Conditional Random Field (CRF) model [17], a conventional Bi-directional long short term memory (Bi-LSTM-CRF) model [18] and a state-of-art Bidirectional Encoder Representations from Transformers (BERT) model [19] were employed, evaluated and compared for this task, from which the highest-performing model was selected for downstream applications.

Aim 2: Develop methods to standardize diverse COVID-19 tests recorded in Electronic Health Records (EHRs)

Using part of the information model from Aim 1 as the annotation axes, we further developed a COVID-19 test normalization tool that can map local test names to the standard LOINC codes. The tool was evaluated using COVID-19 test data from eight healthcare systems and was adopted by other research programs.

Aim 3: Demonstrate the utility of the standardized lab test information by conducting a real-world study of re-infection of COVID-19

Using a large EHR database across the US, we conducted a comprehensive feature analysis of COVID-19 patients with re-positive test results. A hybrid searching strategy,

combining the standardized lab test information and ICD-10 [20] was applied to enhance the completeness of selected patient data. The study investigates the overall cumulative re-positivity rate, followed by a thorough detailed review of 23 patients at high risk of re-infection.

1.2 Literature Review

In this section, we review relevant work on the lab test ontology, lab test extraction, and lab test standardization.

1.2.1 Relevant work on ontology development specifically for lab tests in pandemics

Pandemics since the 20th century

A pandemic usually refers to a global epidemic induced by a 'new' virus to which humans possess little or no pre-existing immunity [3, 21]. Such a virus may be a new virus, such as the human immunodeficiency virus (HIV), which has caused more than 35 million deaths since its outbreak in 1981 [22]; re-emergence of existing pathogens causing fulminant outbreaks of infectious diseases, such as cholera and the plague, hinders the acquired immunity due to their extremely high mortality rate [23, 24]; or a new virus subtype, such as the influenza A/H1N1 and A/H3N2 , which are the pathogens

of the famous 1918 flu and the 1968 flu, causing 17-100 million and 1-4 million deaths worldwide respectively [25, 26]. Similarly, the novel coronavirus 2019 is a new subtype of the coronavirus family, as other subtypes of coronavirus are well-known to be the pathogens of 2003 SARS [27] and 2015 MERS [28], both of which are fatal epidemics. Compared to the SARS and MERS viruses, the COVID-19 virus has a longer incubation period (a period when patients are asymptomatic) and is more likely to spread [29], eventually leading to the severe pandemic.

Lab tests during COVID-19

Since the outbreak of COVID-19, many SARS-CoV-2 tests have appeared on the market due to the lack of a standard protocol setup [30]. According to Ravi et al. [31], the COVID-19 lab test can be classified as diagnostic detection and antibody detection. Diagnostic detection is primarily used for the diagnosis of active COVID-19, which focus on the detection of nucleic acid or viral antigen. Antibody tests are applied to identify the disease by measuring the antibodies produced in the body against SARS-COV-2.

In Feb 2020, after the CDC developed the first laboratory diagnostic test kit to detect SARS-COV-2 nucleic acid, a vast majority of commercially or laboratory developed nucleic acid test kits got approved under the FDA EUA. The current standard for SARS-CoV-2 diagnosis is RT-PCR, which serves to amplify specific related genes [32]. At

present, most RT-PCR assays use oligonucleotide primers and probes, which are selected from different gene regions of SARS-CoV-2 virus, including the envelope (E) [33–35], nucleocapsid (N) [36–38], spike (S) [39] and/or open reading frame 1 ab (ORF1ab) genes [40, 41].

Another major test category is the antibody test, also known as serology test. The main strain antibody tests consist of rapid diagnostic tests (RDTs), such as the lateral flow assay [42], enzyme-linked immunoassays (ELISAs) [43], neutralization assays, and chemiluminescent immunoassays [44, 45]. Lateral Flow Immunoassay, ELISA, and chemiluminescent immunoassays are frequently employed to detect IgG and IgM antibodies. Neutralization assay measures the amount of neutralizing antibodies, which could bind to virus and block its replication.

Currently, both the RT-PCR tests and the antibody tests can be qualitative or semi-quantitative. The specimens applied for lab tests comprise upper respiratory specimens, lower respiratory specimens, and blood products (serum, plasma, or whole blood) [31]. The lab tests can be conducted in laboratories accredited by the Clinical Laboratory Improvement Amendments (CLIA) [46], or point-of-care setting, or even at home [32].

The prices of the reluctant lab test products in the market vary drastically, from $50 to $200/per test [47] for the same kind of test.

Existing efforts on standardizing lab tests

Ontology for Laboratory Test Prescription and Reporting (LABO) - Adrien BARTON et al. [48] developed LABO, an ontology for formalizing lab test prescriptions and reporting documents. It is based on the Open Biological and Biomedical Ontology (OBO) [49]. LABO is a major part of the core ontological model designed to facilitate interoperability between different clinical data sources.

Disease Oriented LOINC Ontologies for Public Health Reporting - Karen Eilbeck et al. [50] developed an ontology to classify the terms used to describe LOINC-coded tests for Chlamydia, and extended it to handle tests for tuberculosis in order to check the scalability of this model. The requirements for tuberculosis laboratory test reporting in Utah and New York City were scrutinized, which were gathered for the CDC's Reportable Conditions Knowledge Management System (RCKMS) project [51]. It furnished the basis for manual queries of LOINC for possible tests for tuberculosis and revealed new terms that could be added to the ontology. For each test, they created a new ontology term with a logical definition and applied the HermiT reasoner [52] to automatically categorize the tests into the ontology terms. The LOINC database supplies a structure conducive to the development of an application ontology, which supports

epidemiologists in the task of managing code sets that meet reporting standards. With the improved ontology, the automated classification strategy can be reproducible and be extended to handle new diseases and problems.

Ontology for LOINC – SNOMED CT Harmonization of Observable Entities - James R. Campbell et al. [53] created an ontology to represent the lab prescription (LOINC) and their ordinal or qualitative results (SNOMED CT) in order to support clinical decision making and research by integrating anatomic and molecular pathological data. This ontology concentrated on pathology reports of the colorectal and invasive breast cancer, which follows the SNOMED CT hierarchy of "Observable entities", "Body structures", "Clinical findings", "Techniques", "Property types", "Situations", "Substances", "Attributes", and "Qualifiers". In this task-specific ontology, the researchers developed a total of 194 new concepts according to the hierarchy. The LOINC terminology was employed as a supplement for wider concept coverage.

The Community-based Ontology for Coronavirus Disease (CIDO) - Coronavirus Infectious Disease Ontology (CIDO) [54] covers multiple areas of coronavirus diseases, including etiology, transmission, epidemiology, pathogenesis, diagnosis, prevention, and treatment. Complying with the OBO Foundry principles, CIDO adopts an extensible ontology development strategy compatible with OBO. Presently, CIDO contains more

than 4,000 terms imported from approximately 20 other ontologies, such as ChEBI [55], Human Phenotype Ontology [56], Disease Ontology [57], and the NCBI taxonomy ontology (NCBITaxon) [58].

1.2.2 Relevant work on medical information extraction

Information extraction (IE) is an area of NLP, implying the automatic extraction of structured entities, concepts, events, relevant attributes, and relations from free text [59–61]. An IE application typically consists of the following subtasks: named entity recognition (NER) which recognizes entity names from the text (e.g., body locations, drugs, etc.) [62]; coreference that links names referring to the same entity [63]; and relation extraction to determine the relations between entities [64].

IE work in the biomedical domain

Wang et al. [65] provides a comprehensive review about clinical applications of IE, collected from 263 publications in the medical domain from 2009 to 2016. According to their research, IE technology is predominantly applied to medical areas such as clinic notes, pathology reports and radiological reports. In terms of the disease categories, cancer and cancer-related domains are among top areas. IE tools, such as cTAKES [66], MetaMap [67], MedLEE [68] are commonly employed in the clinical IE tasks..

Rule-based IE approaches

According to Wang et al [65], a rule is "usually a pattern of properties that need to be fulfilled by a position in the document". A common form of rule is regular expression, using a series of traits to define a search pattern. A clinical information extraction system is often composed of multiple rules, developed through artificial knowledge engineering, utilizing knowledge bases, or a hybrid system. Medical terminologies and knowledge bases, such as Unified Medical Language System (UMLS) [69], SNOMED CT medical terminology [16], or controlled lexicon such as RadLex [70] (for radiology terminology) are often used lexicon sources in such approaches.

The rule-based IE approaches are widely accepted in the applied clinical field because that they usually can achieve high performance on most tasks in this field, although some researchers consider they are obsolete and are more likely to explore how to develop and apply the state -of-art, machine-learning based approaches to solve medical problems [65].

Machine learning-based IE approaches

Recently, the machine-learning based IE approaches are experiencing unprecedented rapid advancement [71]. Most of the traditional approaches train models via supervised learning algorithms, including Support Vector Machine (SVM) [72] (Cortes & Vapnik,

1995) method, CRF [17], and Generalized nearest neighbor (NNge) [73], etc. Now, deep learning methods show significant improvements in many NLP tasks including IE. The prevalent approaches include the original convolutional neural network (CNN) [74] model, recurrent neural network (RNN) [75] model, and their derivative models (e.g. the long short term memory, LSTM model [76]). Since 2019, pre-trained language models based on contextual embeddings (e.g., BERT [19]) have shown significantly improved performance on multiple NLP tasks including in medical IE tasks [77]. Nevertheless, despite the widespread acceptance of deep learning-based approaches in the academic research domain, the sole machine learning-based approaches are still less utilized in the specific task-oriented applied clinical domain than rule-based approaches [65].

1.2.3 Relevant work on lab test standardization

According to CDC [78], the lab test standardization is defined as "Equivalent results, within clinically meaningful limits, among different measurement procedures for the same laboratory test". Standardization is significant for clinical practice and can alleviate the dilemma of commutability among healthcare facilities. For medical laboratories, the International Organization for Standardization (ISO) 15189 (Medical laboratories – Particular requirements for quality and competence) is widely followed [79, 80]. Another commonly used standard when transferring medical data is the Health Level 7 [81],

which is currently collaborating with ISO to establish standards of health data exchange. ISO and HL7 cover a range of areas from the laboratory test design, quality control and clinical information exchange. Another laboratory standard, LOINC [15], was initiated by Clem McDonald in 1994. Different from HL7 and ISO, LOINC focuses on the vocabulary standard for clinical and laboratory observations (e.g. describing tests and measurements).

With the outbreak of COVID-19 in 2020, FDA started to authorize emergent use of novel lab tests for SARS-COV-2 under the EUA system. The importance of standardizing lab test information has caught the attention of both researchers and policy makers. Several consortia have been formed to construct large clinical data networks for COVID-19 research, including the National COVID-19 Cohort Collaborative (N3C) [82], the international EHR-derived COVID-19 Clinical Course Profiles (4CE) [83], etc. The large-scale research consortia working to establish medical data networks are also developing manual or automatic methods to normalize various lab tests from databases of different participants into a common vocabulary.

To facilitate the effort on standardizing COVID-19 lab testing data, LOINC and CDC issued mapping guidelines for the in vitro diagnostic (IVD) kits to assist diagnostic kits developers manually map the test names to LOINC codes [84]. On June 4, 2020, the U.S.

Department of Health & Human Services (HHS) also launched the implementing of Coronavirus Aid, Relief, and Economic Security (CARES) Act (U.S. Department of Health and Human Services (HHS), n.d.) to reinforce that all COVID-19 lab tests conducted at the CLIA [46] accredited sites should be reported to appropriate state or local public health department in a standardized vocabulary.

Chapter 2: Information Extraction of Covid-19 Lab Test Information from EUA Documents

2.1 Introduction

Human history has been marked by severe pandemics, such as cholera [23], plague [24], and smallpox [85] which had extremely high mortality rates and hampered the development of acquired immunity. With the global integration and improved public health interventions, the epidemiology of pandemics in the 20th century, such as the 1918 flu [25] and the 1968 flu [26], changed from severe, high mortality infectious disease to those respiratory diseases that spread easily and rapidly [86]. Similarly, COVID-19 has rapidly spread globally since the emergence of the novel coronavirus in late 2019, becoming the first pandemic of the 21st century. The novel coronavirus of 2019 is a new subtype of the coronavirus family, as other subtypes of coronavirus are well-known to be the pathogens of SARS and MERS, two deadly epidemics. On the one hand, compared with SARS and MERS viruses, the COVID-19 virus has a longer incubation period (a period when patients are asymptomatic) and is more likely to spread [29], eventually resulting in severe pandemics. On the other hand, compared to epidemics such as SARS and MERS, pandemics usually have multiple 'waves' of outbreaks, which prolongs its lifespan and significantly increases the disease burden [87].

Therefore, a robust and expeditious lab testing system plays a particularly crucial role in the fight against the disease in the early stage of a pandemic outbreak. A fast-responding laboratory development-approval-distribution system provides the information necessary for effective public health surveillance and early interventions, which is extremely pivotal in preventing the progression from epidemic to pandemic and in mitigating pandemic outbreak.

Fortunately, advances in biology and other technologies, along with the FDA EUA system, have led to the development of COVID-19 lab tests faster than ever before. In a short period of time, many COVID-19 tests have been developed, authorized, and distributed to the market for diagnosing and monitoring the disease. Table 1 shows the COVID-19 lab test statistics by May 25, 2021, together with the three pandemics since the 20th century, including SARS in 2003, all belonging to the same family of COVID-19 [21, 22, 25–27].

Table 1: Lab test statistics for 4 pandemic/epidemics

Pandemic/Epidemic	No. Tests	Last Outbreak period	Pathogen isolation from the initial outbreak	From the first case to pandemic
Influenza	244	multiple waves	70 years from the 1918	During WW1,

(H1N1/H3N2)		until present	H1N1 flu	unclear
AIDS(HIV1/HIV2)	257	1980s-2000s	5-6 years	More than 10 years
SARS	21	2002-2003	5 months	epidemic
COVID-19	131	2019-current	1 month	2.5 months

The numbers of influenza virus, HIV, and SARS virus detection were extracted from the recently released Logical Observation Identifiers Names and Codes (LOINC), while the number of COVID-19 tests was from the recently released COVID-19 LOINC code sets.

As can be seen in Table 1, the rate of transmission and the development of lab tests for COVID-19 far exceed that of the 2003 SARS epidemic, whose pathogen belongs to the same coronavirus family as COVID-19. Although the absolute amount of COVID-19 lab tests is smaller than for flu and AIDS, the time to identify the disease pathogen for COVID-19 and develop corresponding lab tests is much shorter. Moreover, COVID-19 has the shortest time frame from the first case to pandemic. As global integration and technological improvements accelerate, it is reasonable to expect that future pandemics will occur in a faster and more dangerous manner, thus requiring a faster development of more lab tests.

Under the FDA EUA mechanism, a quick and robust lab test system for new pandemics can facilitate early response to disease, especially for future outbreaks. However, many lab tests details are embedded in narrative text (i.e., FDA EUA documents are free text in .pdf format), making data sharing difficult and affecting the efficiency of medical resource allocation, clinical research, and early disease mitigation.

To address this challenge, LOINC and CDC timely issued the mapping guidelines for in-vitro diagnostic (IVD) kits [84], to assist diagnostic kits developers in manually mapping the test names to LOINC codes. The U.S. Department of Health & Human Services (HHS) also launched the Coronavirus Aid, Relief, and Economic Security (CARES) Acton to reinforce that all COVID-19 lab tests conducted at CLIA certified sites should be reported to the appropriate state or local public health department in a standardized vocabulary. (U.S. Department of Health and Human Services (HHS), n.d.) However, during this public emergency of new pandemic, there remains a gap between the release of lab test standardization policies and the rapid emergence of lab tests, and the gap time may increase when the lab test production capacity enhances in future pandemic outbreaks. To prepare for potential future pandemic outbreaks and the highly possible surge of new lab tests, it is important to establish an automated approach to lab test standardization.

An ontology is a “formalized encoding of knowledge that enables machines and computerized agents to understand domain information” [88]. It also promotes the understanding of the nature of a concept, enabling software agents to ratiocinate from the semantic logical connections between concepts, which is machine-based.

Hence, in this study we adopted the ontology approach to develop an information model that encompasses the key attributes of lab tests emerging from the COVID-19 pandemic, and further developed an automatic entity recognition model to extract such information from EUA documents of COVID-19 lab tests. Specifically, this study contributes to informatics and medicine in two main aspects:

1. Develop a pandemic specific information model for new COVID-19 diagnostic tests, which utilizes concepts from LOINC and SNOMED CT as well as some original concepts.
2. Based on the information model, further develop an information extraction tool to automatically identify the entities for COVID-19 lab tests in FDA EUAs.

2.2 Methods

The workflow of our study includes the following steps: (1) data collection; (2) information model development and corpus annotation; (3) automatic identification of attributes for lab tests.

2.2.1 Data sources

Since all the COVID-19 lab tests in the U.S should be used under the FDA EUA, I collected 378 (until 05/25/2021) EUAs of COVID-19 lab tests from the FDA website for our corpus. The EUAs in pdf format were preprocessed using python to convert them to text files. The most informative part of each EUA (the abstract section or the first paragraph if there is no abstract section) was retained as our corpus. Figure 1 shows an example of each type of EUA. The circled parts were extracted into text files for the NER task.

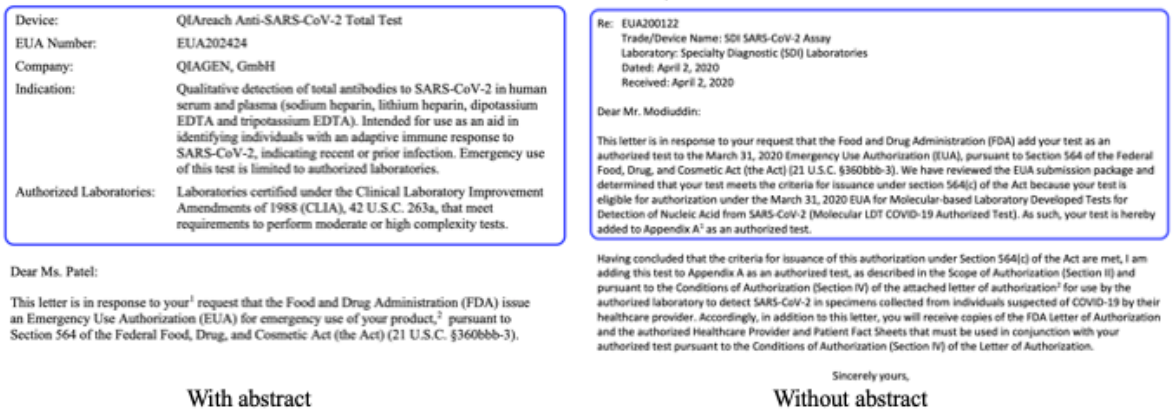

Figure 1: Examples of each type of EUA

2.2.2 Information Model Development

The COVID-19 lab ontology concepts are created by reviewing all the current COVID-19 lab tests under the FDA EUA and integrating LOINC and SNOMED CT terminologies for medical ontologies. The EUA authority allows FDA to permit unapproved medical products or unapproved uses of approved medical products in public health emergencies. The National Library of Medicine (NLM) terminology has also been examined because it furnishes several well-known 'knowledge infrastructure' resources, for example, UMLS Metathesaurus [69], which documents numerous synonyms of biomedical terms and categories of biomedical concepts that facilitate clinical NER. The COVID-19 lab information model is formalized in the Ontology Web Language (OWL) using Protégé

5.5, while the logical consistency and taxonomy in the model are checked using HermiT 1.4.3 reasoner. The information model is evaluated with the assistance of experts in the field of medical informatics, physicians, and epidemiologists. For the subset of information used for the following IE task, the coverage of each entity is checked using the COVID-19 lab test EUAs and the news releases retrieved from the Internet.

2.2.3 Annotation Framework

The subset of the COVID-19 lab ontology applied to the NER task comprises 10 top-level classes of the information model as the annotation axes, which are “TestName”, “Component”, “Scale”, “Method”, “System”, “Facility”, “SamplingMethod”, “Time”, “Institution”, and “SamplingPersonnel”. Among the 10 axes, “Component”, “Scale”, “Method”, “System” and “Time” are imported from LOINC. The definitions of 10 axex are shown below in Table 2 and annotation examples are shown in Figure 2. All annotations are completed using the Clinical Language Annotation, Modeling and Processing (CLAMP) Toolkit [89].

Table 2: Annotation Attributes

Attributes	Definition	Example

TestName	Phrases that describe procedures, panels, and measures to discover or find information about COVID-19	"BinaxNOW COVID-19 Ag Card 2 Home Test", "the LabCorp COVID-19 RT-PCR Test"
Component	The substance or entity being measured or observed	"nucleocapsid protein antigen from SARSCoV-2", "COVID-19 IgG/IgM"
Scale	How the observation value is quantified or expressed	"Qualitative", "Quantitative"
Method	A high-level classification of how the observation was made	"molecular nucleic acid amplification test (NAAT)", "real-time loop mediated amplification reaction"
System	The specimen or thing upon which the observation was made	"fingerstick blood samples", "nasal swab"

Facility	The place where the test can be performed	"at-home use without a prescription", "high and moderate complexity laboratories"
SamplingMethod	In which way the specimen are collected	"Self-collected under observation"
Time	The interval of time over which an observation was made	"twice over three days with at least 36 hours between tests", "15 days after disease onset"
Institution	The place(lab, drug company, universities…) where the test was developed	"Rutgers Clinical Genomics Laboratory", "Yale School of Public Health"
SamplingPersonnel	The person who performed the sample collection	"the healthcare provider", "any individual older than 18 years old"

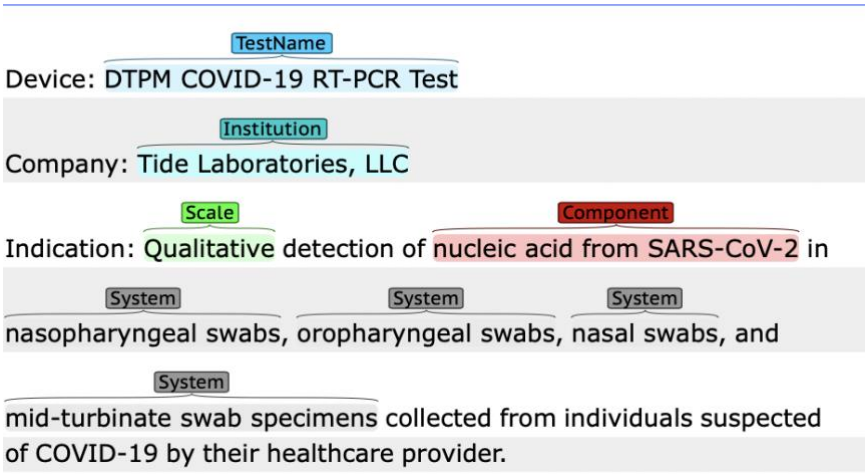

Figure 2: Annotation Examples

2.2.4 Information Extraction Model Development

In this study, a CRF-based [17] NER model was first developed as a baseline using the CLAMP tool. Then we employed two state-of-art machine learning models for this task. The first one is the fine-tuning BERT model proposed by Devlin et al. [19], and the other one is the Bi-Long-Short-Term-Memory model with a CRF layer on top (Bi-LSTM-CRF) [90], which has been proven to perform well in natural language processing tasks. In this task, we substituted BIO tags for the entities in the input header, where "B" indicates the beginning of the entity, "I" represents the subsequent tags in the entity, and "O" denotes other non-entity tags. The text is first processed for sentence boundary detection, tokenization and position-of-sentence (POS) tagging, and then the "BIO" labels are

applied. The output of the Bi-LSTM-CRF model is the predicted probability of "B", "I", "O" for each token, from which the highest is selected. For the BERT model, add a [CLS] token at the beginning of each sentence for classification tasks. The output of the fine-tuning process is the final hidden vector of [CLS] tokens, which represents the semantics of the whole sentence. If an entity is classified into a certain axis, the classification label for this axis is "1", otherwise it is "0". The probability of the classification label is calculated by softmax function [91]. Therefore, the context of such keywords should also be considered.

2.2.5 Experiment and Evaluation

The annotated corpus is randomly divided into training and test sets in a ratio of 4:1. The training set is employed for model development, and the test set is utilized for evaluation.

$$percision = \frac{ture\ positive}{ture\ positive + false\ positive} \quad (1)$$

$$recall = \frac{ture\ positive}{ture\ positive + false\ negative} \quad (2)$$

$$F1 = \frac{2 \times presicion \times recall}{precision + recall} \quad (3)$$

2.3 Results

Figure 3 illustrates the information model developed for COVID-19 lab tests. The evaluation results indicate that the COVID-19 lab information model has a good structure

and concept coverage, and all required modifications were made to optimize the information model concepts structure and coverage, ensuring the adequacy of the semantic categories for each concept. The coverage of the selected 10 core concepts on the EUA dataset is shown in Table 3.

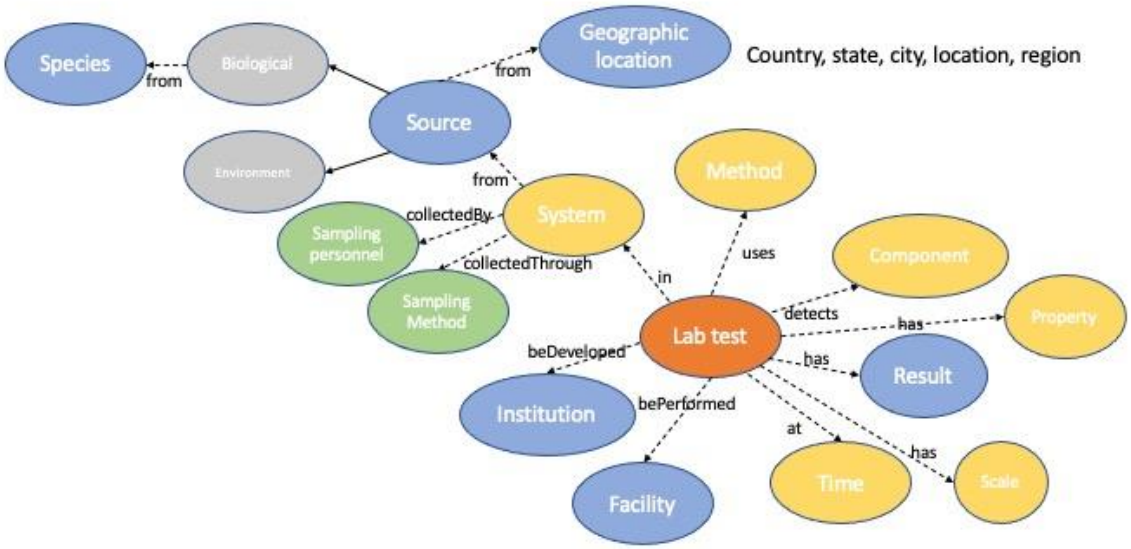

Figure 3: The COVID-19 lab test information model

Table 3: The coverage of the 10 concepts on the EUA dataset

Concept	**Coverage**	**Concept**	**Coverage**

TestName	97%	Method	12%
Component	95%	Facility	98%
System	99%	Time	15%
Scale	90%	SamplingMethod	31%
Institution	85%	SamplingPersonnel	22%

The “Time” concept is less well covered than other concepts because it is only included in the EUAs for some antibody tests and a few tests which support regular screening. Most of the “SamplingMethod” and “SamplingPersonnel” axes are only mentioned in the EUAs for over the counter or point-of-care lab test, or those can be performed in community settings. The concept of “Method” is also referenced in fewer corpora, as some lab tests omit the default PCR method in the abstract/first paragraph of the EUA for which the test is applied for RNA detection.

Figure 4 demonstrates the frequency of words in the training set for the NER task. The most frequent tokens are mainly from the “Component”, “System”, and “Facility” axes. This finding is consistent with the high coverage of these concepts in the training set, and to some extent verifies the homogeneity of the three concepts across the corpus.

Figure 4: Word Cloud of the corpus

Table 4 presents the statistics for the 10 types of entities in the EUA dataset. The frequency of "System" remarkably exceeds other concepts because most lab tests can be conducted on multiple specimens.

Table 4: Concept distribution

Concept	Total No. of Entities	No. of Entities per EUA
TestName	393	1.0397
Component	440	1.1640

System	1609	4.2566
Scale	393	1.0397
Institution	331	0.8757
Method	46	0.1189
Facility	441	1.1667
Time	58	0.1534
SamplingMethod	379	1.0003
SamplingPersonnel	82	0.2179

Table 5 reveals the NER task performance of the three models on all the 378 EUAs. The results indicate that both the BERT and the Bi-LSTM-CRF outperformed the baseline CRF model, and they achieved similar F-1 score.

Table 5: Overall performance of the CRF model, BERT and the BI-LSTM-CRF model

Concept	Precision			Recall			F1		
	CRF	BERT	BI-LSTM-CRF	CRF	BERT	BI-LSTM-CRF	CRF	BERT	BI-LSTM-

									CRF
TestName	90.60%	96.33%	98.50%	90.80%	97.77%	98.30%	90.70%	97.04%	98.40%
Sampling-Method	77.30%	87.29%	93.90%	69.90%	90.38%	94.40%	73.40%	88.81%	94.20%
Sampling-Personnel	72.60%	75.82%	89.50%	54.90%	79.31%	80.00%	62.50%	77.53%	84.50%
Scale	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Time	87.30%	91.53%	91.40%	82.80%	90.00%	91.40%	85.00%	90.76%	91.40%
Institution	92.80%	98.32%	95.60%	85.20%	98.32%	95.90%	88.80%	98.32%	95.80%
Facility	88.40%	95.09%	89.60%	85.00%	97.19%	89.40%	86.70%	96.13%	89.50%
Method	87.80%	96.33%	95.70%	78.30%	95.65%	95.70%	82.80%	95.65%	95.70%
Component	97.20%	97.39%	99.30%	95.00%	98.24%	99.10%	96.10%	97.81%	99.20%
System	96.10%	98.43%	98.90%	95.20%	98.07%	98.90%	95.60%	98.25%	98.90%
Overall	92.84%	96.33%	97.00%	89.88%	96.96%	96.80%	91.34%	96.62%	96.90%

2.4 Discussion

In this study, we developed and evaluated a COVID-19 lab test information model according to the ontology technology. Subsequently, applying the subset derived from the information model as entity concepts, we developed, evaluated, and compared three machine learning-based NER pipelines to automatically extract the COVID-19 lab tests from the FDA EUAs documents.

From the performance comparisons we found that the baseline-CRF achieved a high overall performance (91.34%), with 92.84% precision and 89.88% recall. Entity types that are highly consistent across the corpus (such as Scale, Component, and System) perform much better than types with more variety. Both the BERT and the Bi-LSTM-CRF models significantly enhance the performance in terms of precision and recall, resulting in improved the F-1 score. The BERT model is a state-of-are algorithm that has achieved record-breaking performance in multiple NLP tasks. The Bi-LSTM-CRF is a traditional recurrent neural-network algorithm that has been proven to perform excellently in NER tasks. In our study, the Bi-LSTM-CRF model achieved a slightly higher performance than BERT on “SamplingMethod” and “SamplingPersonnel”, two concepts that occur less frequently in our corpus than the others. This result is comparable with Ezen-Can’s finding [94], suggesting that small corpus size may affect the performance of the BERT model. It is also possible that more fine-tuning is desired for the BERT model to reach a better F-1 score. On the concepts with long sentences,

such as “Institution” and “Facility”, the BERT model significantly outperformed the baseline-CRF model and the Bi-LSTM-CRF model, exhibiting its merit in handling long sentences.

From the case of COVID-19, we can discern that the development and approval of lab tests during a pandemic undergoes a unique process. In the U.S., all the lab tests for COVID-19 enter the market through the FDA’s EUA rather than the traditional approval process, leading to a rapid response to the public health emergency. The formal format with highly homogeneous language used in the EUA also facilitates the adoption of NLP tools to automatically extract information from the EUAs documents. By automatically retrieving standardized lab test information from the EUAs, it can assist clinical healthcare providers, researchers and public agencies in effectively collecting and sharing standardized lab test data in the very early stage of a pandemic outbreak, thereby improving the efficiency of healthcare resource allocation and medical research.

To the best of our knowledge, this is the first attempt to develop an ontology-based automatic information extraction tool to consistently represent and identify rapidly emerging lab tests from EUAs approvals. The developed information model can serve as a template and be extended to new lab tests for future pandemics. The ontology based NER model can efficiently retrieve standardized lab test information from EUAs,

demonstrating the implementation significance of NLP technology in response to public health emergencies.

The limitation of our study is that COVID-19 is the only use case for developing the information model and the NER tool, although our ultimate goal is to represent lab tests for any pandemic. Despite the fact that we considered other previous pandemics in the development of the ontology, the lack of EUA documents for those previous pandemic lab tests hinders us from a holistic picture of lab testing in those pandemics. Our future work is to collaborate with domain experts to further test and optimize the ontology, entitling it higher potential to be generalized to future pandemics.

2.5 Conclusion

In this study, we take COVID-19 lab tests as a use case to design an information model for lab tests, as well as machine learning-based NER methods to automatically extract information from the FDA EUAs documents. Our results indicate that the BERT and the Bi-LSTM-CRF model achieves an outstanding performance in the NER task, which demonstrates the application of AI approaches in data collection and standardization in pandemics.

Chapter 3: Covid-19 Testnorm - A Tool to Normalize Covid-19 Testing Names to LOINC Codes

3.1 Introduction

In the last chapter, we developed an information model and a machine learning based information extraction tool to extract structured lab test information from the FDA EUAs for the COVID-19. In this chapter, we will discuss the normalization of COVID-19 lab test information from another data source.

COVID-19 patients' clinical data stored in EHR is an important data source for COVID-19 research. Several consortia have been formed to construct large clinical data networks for COVID-19 research, including The National COVID-19 Cohort Collaborative (N3C) [82], the international EHR-derived COVID-19 Clinical Course Profiles (4CE) [83] and many others.

To efficiently conduct clinical studies across different institutions within a network, one requirement is to normalize clinical data to common data models (CDM) and standard terminologies. One such example is the Observational Medical Outcomes Partnership (OMOP) CDM maintained by the Observational Health Data Science and Informatics (OHDSI) consortium [95]. Among different types of clinical data, COVID-19 diagnostic

tests are critical for all the following analyses, as they are the primary means to identify the confirmed COVID-19 cases. To address the urgency of the pandemic, individual institutions have created local names and local codes for those new COVID-19 tests in their EHRs. Meanwhile, the LOINC has responded quickly by developing a new set of standard codes for COVID-19 tests [84] to guide standard coding of these tests in clinical settings. Nevertheless, there is a lack of mappings between local COVID-19 test names and standard LOINC codes, which hampers cross-institutional studies that rely on normalized clinical data at each institution. Existing natural language processing (NLP) systems such as MetaMap [67] or CLAMP [89] provide concept mapping functions, but none of them has been updated to accommodate new concepts for COVID-19 tests.

To address this urgent need for reliable mappings, we developed an automated tool -- COVID-19 TestNorm -- to normalize a local COVID-19 test name to a standard LOINC code. This tool is available to the community via an open-source package at GitHub and via an online web application. We believe COVID-19 TestNorm can be a useful tool for the secondary use of EHRs for research studies on the pandemic.

3.2 Methods

Using COVID-19 test data collected from eight healthcare systems, we developed a rule-based system to automatically normalize a local test name to a LOINC code for COVID-19. Figure 5 shows an overview of the modules of the COVID-19 TestNorm system, mainly including entity recognition and LOINC mapping modules, with inputs from knowledge components such as lexicons and coding rules. The input lab test names are tokenized first, then specific entities are recognized, and appropriate LOINC codes are automatically mapped based on the coding rules.

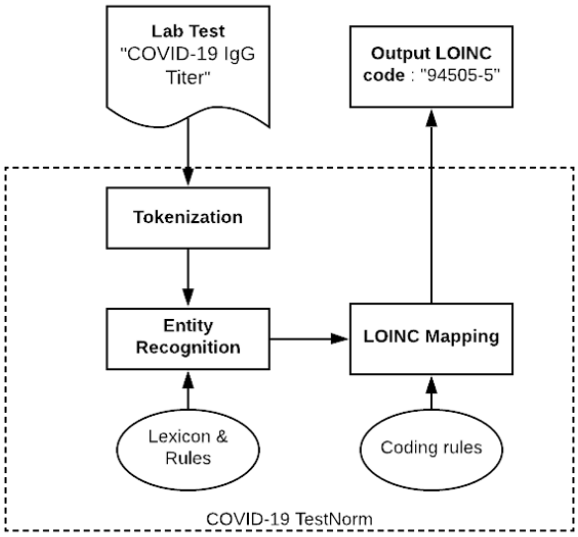

Figure 5: An overview of the COVID-19 TestNorm system

3.2.1 Dataset

We collected COVID-19 test data from eight healthcare systems across the United States, including University of Texas Physicians, Memorial Hermann Health System, University of California San Diego, Mayo Clinic, University of Florida, University of Minnesota, Columbia University Medical Center, and the national Department of Veterans Affairs (collected from 170 medical centers and 1,063 outpatient sites) in April 2020. Data from each institution primarily contained test names, as well as other fields available in local lab tables, such as specimen information. In total, 568 records were collected from the eight sources. Although some institutions provided LOINC codes with the names, we manually reviewed all the records and assigned corresponding LOINC codes. Two annotators followed the LOINC COVID-19 coding guideline5 and manually mapped the 568 records to LOINC codes. The Cohen's Kappa agreement8 between the two annotators was 99.3%. We then randomly divided the dataset into a development dataset (454 records) and a test dataset (114 records). The COVID-19 TestNorm tool was developed using the development dataset and evaluated on the test dataset.

3.2.2 Entity recognition

LOINC describes each concept using six primary axes: Component, System, Method, Time, Property, and Scale9, some of which were included in our COVID-19 entity categories. Our five root categories were Component, System, Method,

Quantitative/Qualitative, which defines if a test returns a qualitative or quantitative result, and Institution, which specifies the manufacturer of the test kit. The LOINC team at Regenstrief has worked with several in vitro diagnostics (IVD) test kits manufacturers and commercial labs to develop and assign appropriate LOINC codes for their SARS-CoV-2 tests. Some of these mappings are listed on the LOINC website [84].

Furthermore, from the manual review of the training set data and coding rules by LOINC [84], we identified that accurate mapping requires more specific values under each root category. For example, for System, which refers to the test specimen, "Serum or plasma", "Saliva", "Nasopharyngeal specimen", "ANY respiratory specimen", and "Unspecified specimen" will lead to different LOINC codes, since the corresponding test methods may vary. In this case, these subcategories of the root category System are essential elements for accurate mapping. This finding also applies to the other root categories. As a result, we divided the five root axes into subcategories. Table 6 lists all the detailed entity categories used in our LOINC coding system, as well as corresponding examples. Once entity categories were defined, we further analyzed the development dataset and manually extracted all related terms for each category, which were appended to the lexicon file used for the COVID-19 TestNorm tool. The lexicon file is publicly available together with the COVID-19 TestNorm software package. Potential users can

manually revise the lexicon file to further improve COVID-19 TestNorm's performance on their local data.

The entity recognition consists of two steps: (a) an initial step that combines dictionary-lookup and regular-expression matching, (b) a disambiguation step that converts the ambiguous tags from the initial step into the final tags according to a set of predefined rules. During the initial step, most information can be captured and tagged to its corresponding category, whereas some ambiguous words need to be further reviewed. For example, the word "IA" can be either mapped to a "method" which represents the abbreviation of "immunoassay" or to a "system" which represents the state "Iowa". We developed context-based rules to determine the correct semantic categories for those terms.

Table 6: Semantic categories used by COVID-19 TestNorm

LOINC axes	**Fine Entity Types**	**Example Values**
Component	Covid19	"COVID-19", "SARS-COV-2"
	Covid19_Related	"SARS-related CoV", "SARS-like CoV"
	RNA_Comp	"RNA", "N gene", "RdRp gene"

	Sequence_Comp	“Whole genome”
	Antigen_Comp	“Ag”, “Antigen”
	Growth_Comp	“Organism”
	Antibody_Comp	“Ab”, “Antibody”, “IgM”, “IgG”
	Interpretation_Comp	“Interpretation”, “Recent infection”
System	Blood	“Blood”, “Serum”, “Plasma”
	Respiratory	“NARES”, “NASAL MUCUS”
	NP	“NP”, "Swab", “NASOPHARYNX”
	Saliva	“SALIVA”, “ORAL FLUID”
	Other	“UNSPECIFIED”, “UNKNOWN SPECIMEN”
Method	RNA_Method	“Non-probe-based”, “NAA”, “PCR”
	Sequence_Method	“Sequencing”
	Antigen_Method	“Rapid IA”, “Immunoassay”, “IA”
	Growth_Method	“Organism specific culture”
	Antibody_Method	“Rapid IA”, “Immunoassay”, “IA”

	Panel_Method	“Panel”, “Panl”
Quantitative_Qualitative	Quantitative	“Cycle Threshold”, “viral load”
	Qualitative	“Presence”, “Ord”
Institution	Manufacturer	“Abbott”

3.2.3 LOINC mapping

LOINC guidelines for COVID-19 tests [84] (as of May 30th, 2020) were followed to guide the development of the initial coding rules, which consist of decision-making algorithms based on extracted entities in the previous step. The coding rules were then iteratively updated using the development dataset collected across institutions. Figure 6 shows the overall decision workflow based on the coding rules (as of May 30th, 2020). It starts with checking manufacturer information, as specific LOINC codes are assigned to known test kits by specific manufacturers. If no specific manufacturer information is available, the tool continues the mapping procedure using test purpose rules. Five test purpose rules are defined based on the tagged entities for Component, Method, and System with the following information: (a) RNA; (b) Sequence; (c) Antigen; (d) Growth; (e) Antibodies. For each test purpose rule, specific tagged entities for the analyte

(Component), specimen (System), Method, and/or Qualitative/Quantitative are further checked to map to appropriate LOINC codes.

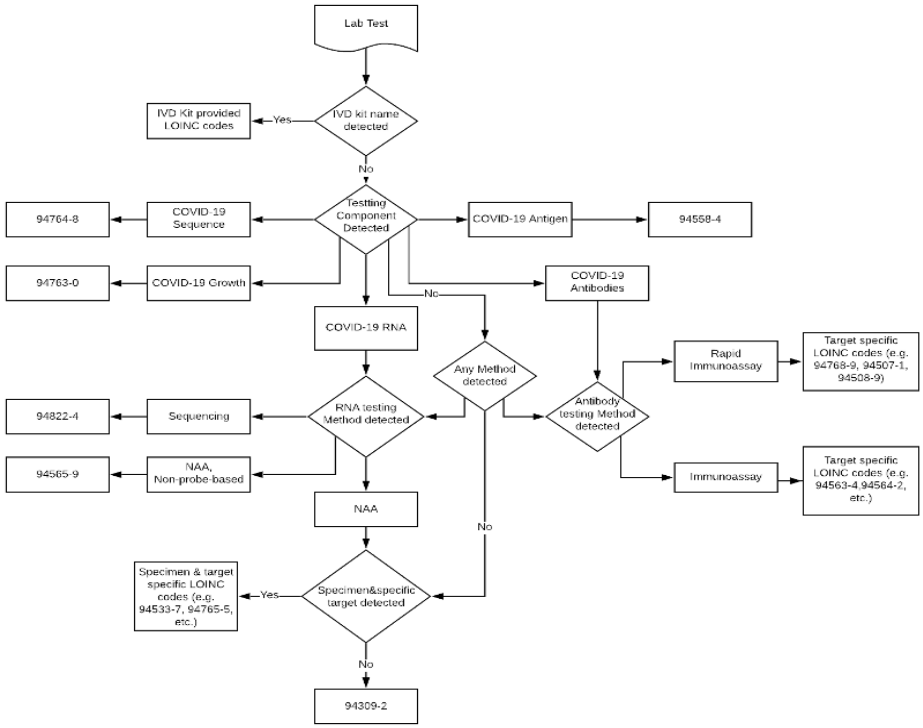

Figure 6: Coding rules for LOINC mapping

IVD: in vitro diagnostics, NAA: nucleic acid amplification

3.2.4 Evaluation

We developed the COVID-19 TestNorm tool using the development set (454 records) and evaluated its performance using the independent test set (114 records). We compared the system's output with the manually annotated gold standard and reported the accuracy

of the system (the percentage of correct LOINC codes generated by the system among 114 records).

3.3 Results

Table 7 shows the distribution of different COVID-19 tests' LOINC codes on the full annotated dataset (568 records). LOINC codes 94759-8 ("SARS-CoV-2 (COVID19) RNA [Presence] in Nasopharynx by NAA with probe detection"), 94500-6 ("SARS-CoV-2 (COVID19) RNA [Presence] in Respiratory specimen by NAA with probe detection", and 94309-2 ("SARS-CoV-2 (COVID19) RNA [Presence] in Unspecified specimen by NAA with probe detection"), were the most frequent codes across institutions, of which "94759-2" is the most frequent one with over 40% of occurrences in the collected dataset. All three codes represent test for SARS-CoV-2 RNA using nucleic acid (RNA) amplification with a probe-based detection method without specifying the gene or region being tested. The 94500-6 code is used for tests that can be run on a variety of respiratory specimens, 94759-8 is specific for nasopharyngeal specimens, and 94309-2 is for unspecified specimens. Nucleic acid amplification with probe-based detection is the most widely used test method so far across the eight sources.

Table 7: Distribution of mapped LOINC codes

LOINC Codes	Total #	Percentage	LOINC Long Common Name
Molecular			
94759-8	240	42.25%	SARS-CoV-2 (COVID19) RNA [Presence] in Nasopharynx by NAA with probe detection
94500-6	202	35.56%	SARS-CoV-2 (COVID19) RNA [Presence] in Respiratory specimen by NAA with probe detection
94309-2	75	13.20%	SARS-CoV-2 (COVID19) RNA [Presence] in Unspecified specimen by NAA with probe detection
94502-2	13	2.29%	SARS-related coronavirus RNA [Presence] in Respiratory specimen by NAA with probe detection
94660-8	11	1.94%	SARS-CoV-2 (COVID19) RNA [Presence] in Serum or Plasma by NAA with probe detection
Antibody			
94563-4	10	1.76%	SARS-CoV-2 (COVID19) IgG Ab [Presence] in Serum or Plasma by Immunoassay

94564-2	4	0.70%	SARS-CoV-2 (COVID19) IgM Ab [Presence] in Serum or Plasma by Immunoassay
94762-2	2	0.35%	SARS-CoV-2 (COVID19) Ab [Presence] in Serum or Plasma by Immunoassay
94504-8	2	0.35%	SARS-CoV-2 (COVID19) Ab panel - Serum or Plasma by Immunoassay
94505-5	2	0.35%	SARS-CoV-2 (COVID19) IgG Ab [Units/volume] in Serum or Plasma by Immunoassay
94507-1	1	0.18%	SARS-CoV-2 (COVID19) IgG Ab [Presence] in Serum, Plasma or Blood by Rapid immunoassay
94508-9	1	0.18%	SARS-CoV-2 (COVID19) IgM Ab [Presence] in Serum, Plasma or Blood by Rapid immunoassay
Other			
56831-1	4	0.70%	Problem associated signs and symptoms
90101-7	1	0.18%	Internal control result

In addition, we also counted the number of unique COVID-19 test codes at each participating site. As shown in Figure 7, the number of unique tests at each site varied,

with Columbia University Medical Center at the top, probably indicating that many test methods have been used in this medical center in New York City.

The overall accuracy of COVID-19 TestNorm on the development set was 98.9%. When evaluated using the independent test set, the system achieved an accuracy of 97.4%, indicating that the rule-based approach was effective in normalizing COVID-19 test names to LOINC codes.

The source code of the LOINC TestNorm tool is available at a GitHub repository [96]. An online web application (https://clamp.uth.edu/covid/loinc.php) is also provided so that users can enter local COVID-19 test names and retrieve mapped LOINC codes automatically.

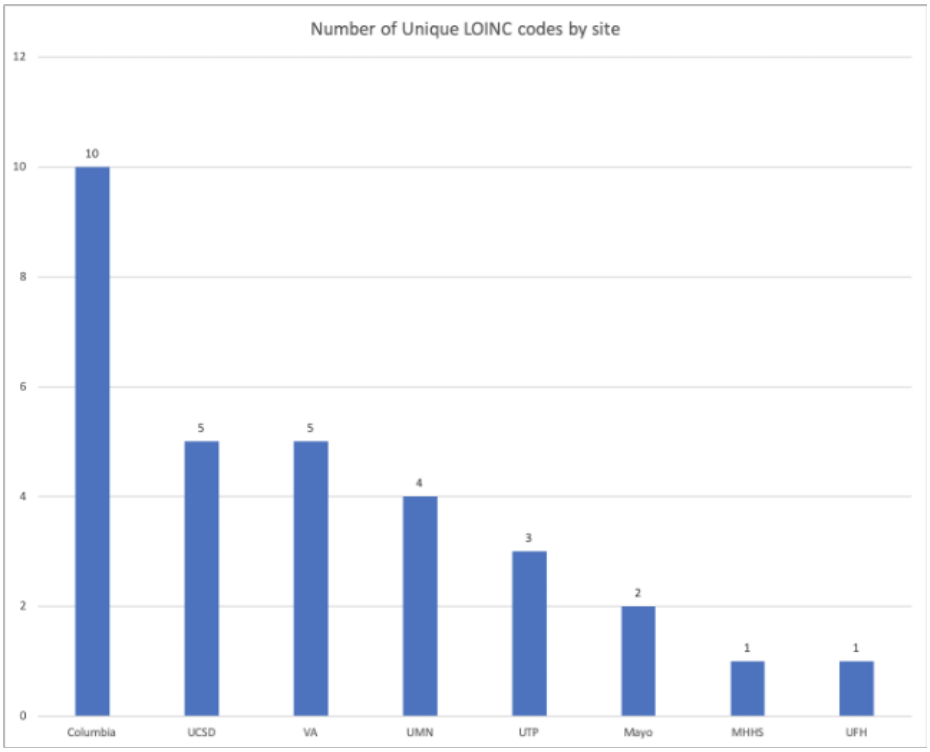

Figure 7: Number of unique LOINC codes by site

UCSD: University of California San Diego, VA: Veteran's Health Affairs, UMN: University of Minnesota, UTP: University of Texas Physicians, MHHS: Memorial Hermann Health System, UFH: University of Florida Health.

3.4 Discussion

In this study, we collected the lab tests from eight healthcare systems across the country. We developed a simple but effective normalization system for mapping COVID-19 lab tests to LOINC codes to facilitate rapid research response to the pandemic. The tool is

publicly available with source code. For ease of use, we developed a web application so that end users can easily map their local COVID-19 lab test names to standardized LOINC codes using the online form, thus improving the efficiency of multi-center data aggregation and global knowledge sharing.

We conducted an error analysis for the mis-mapped codes. TestNorm achieved 100% accuracy on most of the LOINC codes in the test set, except for codes 94500-6 (2 records) and 56831-1 (1 record). For the two errors for 94500-6, one test name was "UF BKR QUEST OVERALL RESULTS LAB17003," and the other was "CONFIRMATORY TEST-QUEST". Both were missed because they do not contain the key entity of COVID-19, which is required by our current coding rules. In the future, we may lift this constraint if we assume that all test names are about COVID-19. For code 56831-1, the original local test name "PATIENT SYMPTOM (SARS COV 2)" does not contain any specific test information, and COVID-19 TestNorm assigned 94309-2 even though the original data came with a specific LOINC code 56831-1, probably due to additional information available to the local hospital only.

LOINC codes are designed for use in clinical settings, assuming all information is available. For secondary use scenarios, data submitted by local healthcare facilities do not always contain such detailed information. When the information is incomplete, more

general LOINC codes will have to be assigned. For example, when the specimen is unknown, LOINC 94309-2 ("SARS-CoV-2 (COVID19) RNA [Presence] in Unspecified specimen by NAA with probe detection") will be mapped, which accounts for 13.20% (75/568) in our dataset.

One of the limitations of this study is that, even though we collected data from eight large healthcare systems across the United States, the sample size and data heterogeneity could still be limited. For example, all codes in our dataset are about molecular and antibody tests. With new tests available in the market, the LOINC code sets for COVID-19 are evolving, i.e., with weekly updates from Regenstrief, as well as continuous updates from the CDC, which maintains a file containing recommended LOINC mappings for test kits currently approved by the FDA (https://www.cdc.gov/csels/dls/sars-cov-2-livd-codes.html). Therefore, it is critical for us to keep updating our tool with new code sets and updated coding rules. When large and diverse samples are accumulated, we will also look into more sophisticated machine learning approaches for this task.

Although we primarily designed COVID-19 TestNorm for secondary use of EHRs for research purposes, the tool could be useful at clinical operational settings or public health agencies as well. Unlike large academic medical centers included in this study, many community hospitals, federally qualified health centers, non-academic medical centers,

and clinics are much less familiar with the difficulties in harmonizing data across multiple systems. Given that HHS has just announced more standard reporting for lab test of COVID-1911, COVID-19 TestNorm could be a handy tool for improving COVID-19 lab reporting quality for both healthcare providers and public health agencies.

3.5 Conclusion

Multi-site data aggregation and normalization are essential for rapid response to COVID-19 research using clinical data. We developed an automated tool to normalize local COVID-19 test names to standard LOINC codes. This offers a foundational first step in enabling test data interoperability for research related to COVID-19.

Chapter 4: Comprehensive Characterization of Covid-19 Patients with Test Re-Positivity in A Large EHR System Across the US

4.1 Introduction Literature Review

In Chapter II and Chapter III, we focused on standardization and normalization of two major types of unstructured COVID-19 lab test data using medical informatics technologies. And in this chapter, a real-world study was conducted, and structured lab test information was employed during data preparation to improve the efficiency and data quality of the study.

A reverse transcriptase polymerase chain reaction (RT-PCR) test is considered the gold standard for detection of SARS-CoV-2 in upper and lower respiratory specimens and for diagnosis of COVID-19. While neutralizing antibodies are detectable for several months following recovery from SARS-CoV-2 infection [97, 98], it remains unknown whether and for how long these antibody responses protect patients from re-infection. There have been many case reports of patients with a second positive PCR test after their PCR results turned negative and symptoms resolved [99]. Most of these are suspected cases of re-infection based on limited clinical or test data; in a minority of suspected cases of reinfection, the viral genome sequences were analyzed and shown to be distinct, strongly supporting a re-infection rather than failure to clear an initial infection [21]. In the

absence of genomic evaluations, the presence of two positive molecular tests separated by negative tests, prolonged time, and clinical resolution of symptoms remains the best surrogate measurement of possible re-infection. Using the Centers for Disease Control and Prevention Common Investigation Protocol for Investigating Suspected SARS-CoV-2 Reinfection [100] as a guide, we conducted a comprehensive evaluation of patients who had repeated positive SARS-CoV-2 PCR tests in a large US COVID-19 electronic health record (EHR) database. We characterize their demographic and clinical characteristics, including their SARS-CoV-2 test journey, symptoms, medication use and COVID-19 related complications.

4.2 Patients and Methods

This retrospective study used the Optum® COVID-19 dataset [101], which implements a low-latency data acquisition model that aggregates de-identified EHR data from providers across the continuum of care.

To achieve the best coverage of all the eligible patients, we employed a hybrid data querying strategy. We first identified the COVID-19 patients using the ICD10 code (U07.1), and then extracted their COVID-19 lab information using both the COVID-19 related LOINC codes and a standardized COVID-19 lab test concept set that we

developed in a previous study (described in Chapter 2 and Chapter 3). As of August 20, 2020, the Optum® COVID-19 dataset included 73,702 patients with a COVID-19 diagnosis code that was laboratory-confirmed with a positive SARS-CoV-2 PCR test, of whom we identified 690 having two positive PCR test results separated by at least one negative test result. The study sample was further restricted to patients who had two consecutive negative test results >24 hours apart between two positive test results (N=79); of these, 4 patients had at least 90 days between their two positive tests and another 19 had at least 60 days between their two positive tests and had accessible demographic and clinical data (Figure 8). If a negative test and a positive test were returned on the same day (<24h), both tests were disregarded.

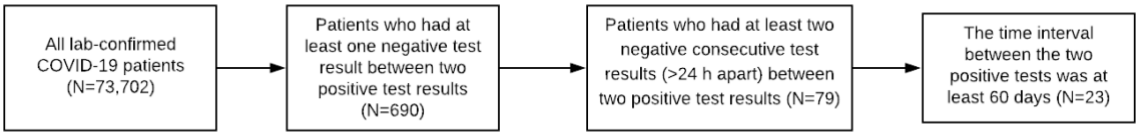

Figure 8: Patient selection flowchart

Demographic and clinical information, including age, gender, race/ethnicity, smoking status, and body mass index (BMI), was extracted. Smoking and BMI were based on the patient's most recent record within one year prior to the index date (first SARS-CoV-2 positive test date).

To get an overall picture about potential re-positivity factors, we created Kaplan-Meier cumulative incidence plots for re-positivity by calculating the cumulative incidence of re-positivity among a subgroup of patients (N=8,618) who had been followed for at least one day since the first negative test after COVID-19 lab confirmation. We also looked at the cumulative incidence rate by age, gender, race, and BMI.

Available EHR data for the 23 patients were manually reviewed. Prevalence of chronic medical conditions which are considered risk factors for COVID-19 was ascertained, including: insulin-dependent type 2 diabetes, hypertension, chronic kidney disease (CKD), respiratory disease (including chronic obstructive pulmonary disease), cardiovascular disease (CVD), atrial fibrillation and immune compromising conditions (including end-stage renal disease on dialysis, HIV, cirrhosis including alcohol-related, solid organ transplant, cancer, and protein-calorie malnutrition). Symptoms typical of COVID-19 were ascertained for each patient during each of two time periods, within 30 days before and after the index date and the second positive test date. In addition, severe clinical outcomes related to COVID-19 illness and medications commonly used to treat COVID-19 were ascertained during each time period, including the following: hospitalization, intensive care unit (ICU) admission, mechanical ventilation, tracheostomy, amputation, or death (at second positive test).

Continuous variables were expressed as median (25th, 75th percentile), and categorical variables as counts (percentages). Missing data were not imputed.

4.3 Results

Figure 9 shows the Kaplan-Meier cumulative incidence of overall re-positivity among the patients (N=8,618) who had been followed for at least one day since the first negative test after COVID-19 lab confirmation. Figure 10-13 shows the cumulative incidence of re-positivity by age, gender, race and ethnicity, and BMI. Most of the re-positivity occurs within 100 days from the negative test, with a cumulative incidence risk at day 100 of approximately 0.12. Men and older individuals have higher re-positivity risk than women and younger patients. Compared to patients with normal or lower BMI, patients with higher BMI had a lower risk of re-positivity. Although Asians had a lower re-positivity risk curve, no clear conclusion can be reached since the sample size of Asian patients is small (N=272).

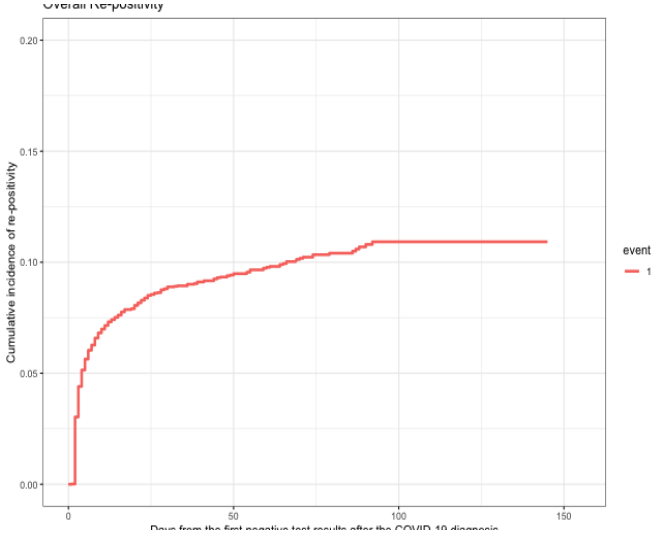

Figure 9: Overall cumulative incidence of re-positivity

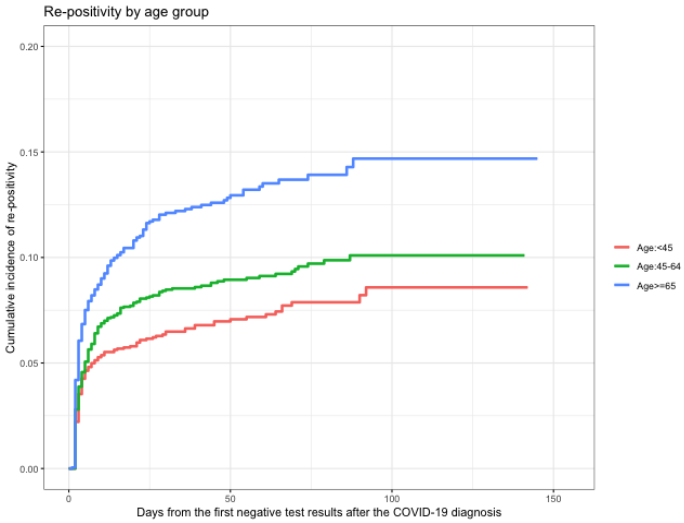

Figure 10: The cumulative incidence of re-positivity by age

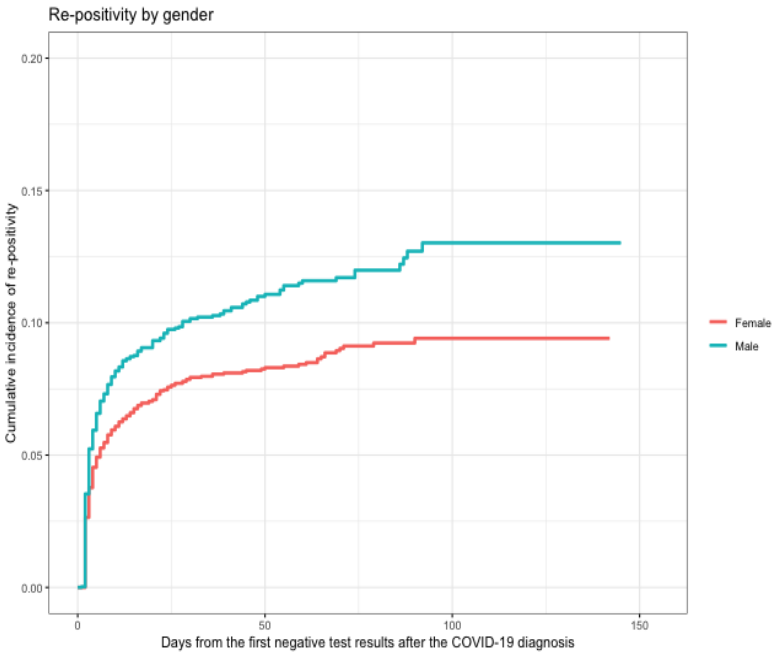

Figure 11: The cumulative incidence of re-positivity by gender

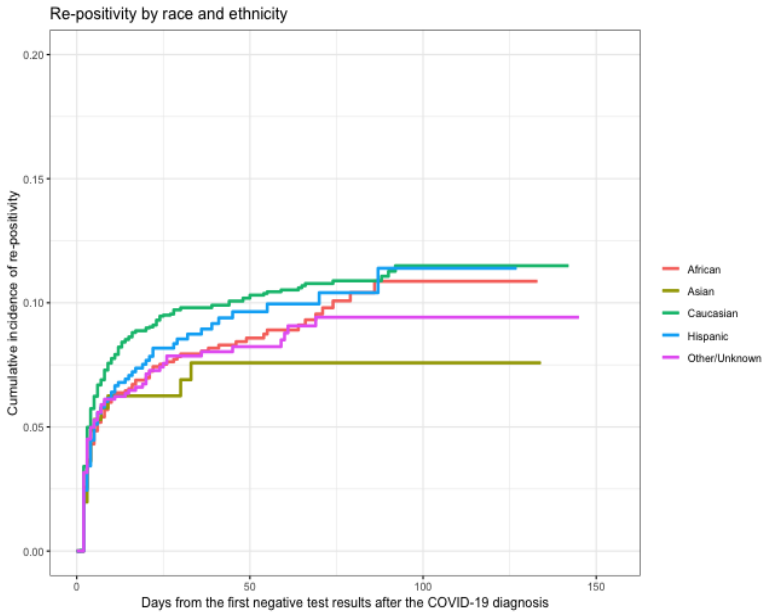

Figure 12: The cumulative incidence of re-positivity by race and ethnicity

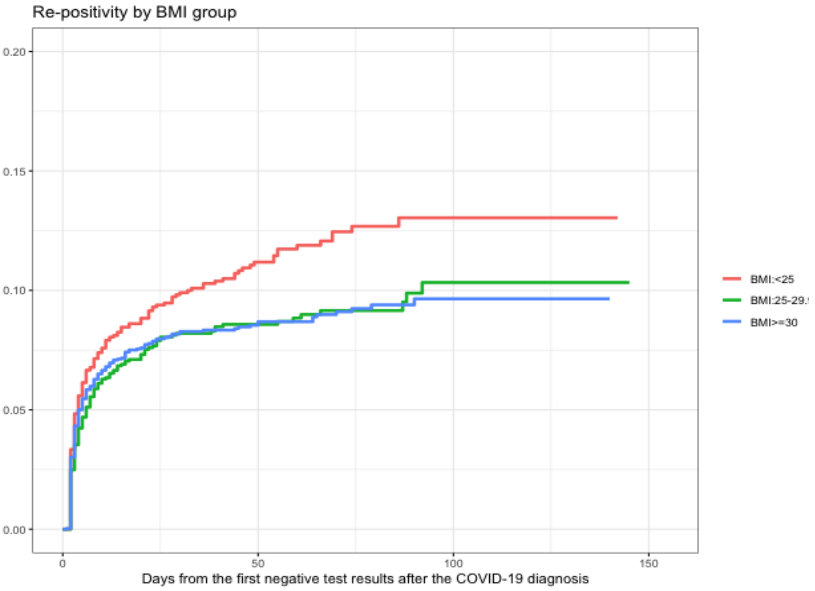

Figure 13: The cumulative incidence of re-positivity by body mass index (BMI) group

For the four patients with at least 90 days between positive tests, the median interval between the two positive tests, separated by two or more consecutive negative tests 24 hours apart, was 100 days (25th, 75th percentile: 96, 107), and the median interval between the first positive and first negative test was 22 days (9, 37) (Figure 14). For the 19 patients with 60-89 days between positive tests, the corresponding intervals were 76 days (25th, 75th percentile: 69, 78) and 32 days (19, 49), respectively (Figure 14).

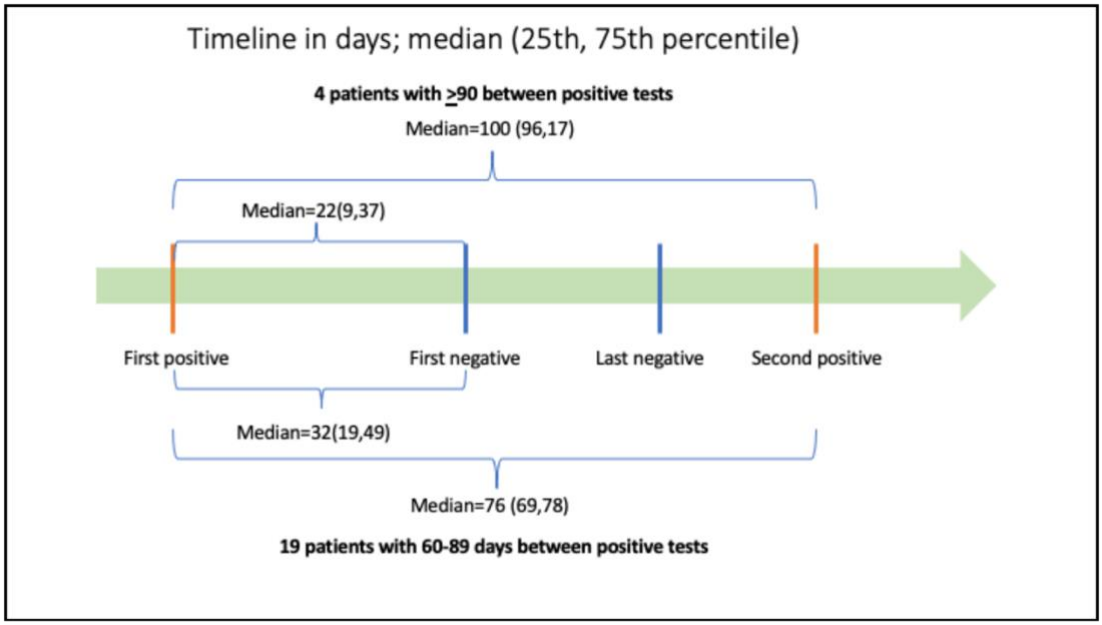

Figure 14: SARS-CoV-2 PCR test timeline (days) for 23 repeatedly positive patients

Median age of the 23 repeatedly positive patients at the index date was 64.5 years (25th, 75th: 53.5, 69.8). Seventeen patients were diagnosed in the Northeast, five in the

Midwest and one in the South; 40% of patients were Black, 40% white, and 20% other/unknown race, 83% had non-Hispanic ethnicity, and 39% were female. Almost 83% smoked within the prior year, and 61% were overweight or obese.

Comorbidity diagnoses and symptom prevalence for the 23 individual patients at the time of each positive test are presented in Table 7, and their PCR test and clinical journeys are shown in Figure 17. Chronic disease prevalence was high, including hypertension (70%), CVD, atrial fibrillation or CKD (each 26%), and insulin-dependent type 2 diabetes or history of venous thromboembolism/long-term anticoagulation (each 22%). Overall, 96% of patients had >2 comorbidities. Most notably, 19 of the patients (83%) had immunocompromising conditions, including two of the four patients with >90 days between positive tests (PT14 and PT19).

For individuals with 45-89 days between positive SARS-CoV-2 tests, CDC investigative criteria include having "a symptomatic second episode and no obvious alternate etiology for COVID-19–like symptoms OR close contact with a person known to have laboratory-confirmed COVID-19." Among the 19 patients in our study with 60-89 days between positive tests, 17 (89%) exhibited symptoms or clinical manifestations indicative of COVID-19 at the time of the second positive test, including 9 (47%) with acute respiratory failure, 8 (42%) with acute kidney failure, 6 (32%) with shortness of breath, 5

(26%) with fever, and 3 with acute embolism and thrombosis (16%). Fourteen of the 19 (74%) were hospitalized at the second positive test, all but four of whom were also hospitalized at the first positive COVID-19 test. One patient was treated with tocilizumab (PT7, at the time of first positive test and during an extended hospitalization) and 4 were treated with dexamethasone after the first diagnosis of COVID-19.

As shown in Figure 15 and Table 8, four of the patients (PT5, PT7, PT12, and PT17) with immune compromising conditions had severe symptoms and lengthy hospitalizations (including ICU and mechanical ventilation for PT12 and PT17) beginning at the first positive COVID-19 test and numerous negative tests, often with their second positive test in close proximity to one or multiple negative tests. Additionally, PT10 had no COVID-19-like symptoms or related treatments at the second positive test, and PT18 had esophageal cancer and no COVID-19-like symptoms at the time of either positive test. The clinical journeys of these six repeatedly positive patients cast doubt about the accuracy of categorizing them as true re-infections.

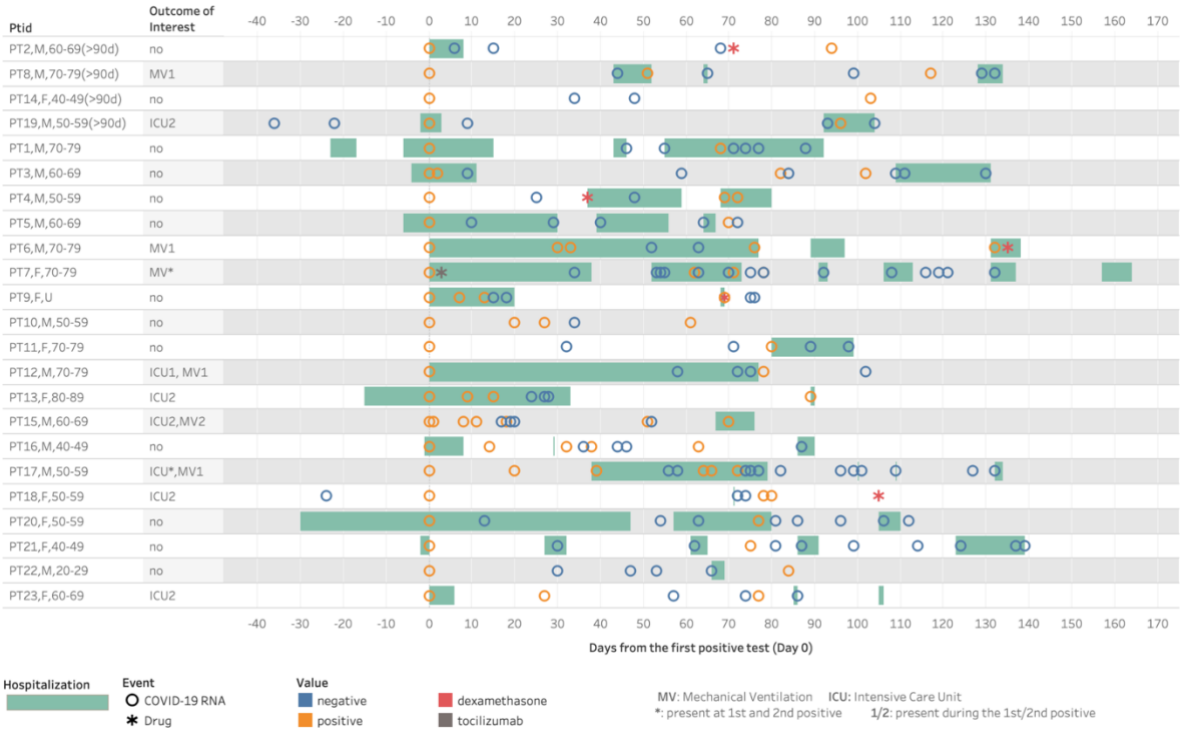

Figure 15: COVID-19 RT-PCR test and clinical journey for 23 patients with repeatedly positive tests

Of the four patients (Figure 15, top) who had >90 days between two positive tests, the record of one immunocompromised patient (PT14) suggests mild-to-moderate disease with few symptoms following both COVID-19 diagnoses. PT19, also immune compromised, had a brief hospitalization at the first diagnosis, followed by ICU admission at the second diagnosis three months later. PT2 had severe symptoms and

hospitalization and treatment with dexamethasone after the first positive test, but no symptoms or treatment at the second positive test.

No patients had cardiac arrest, tracheostomy, amputation, or death.

Table 8: Detailed information for each patient

PT	1	2	3	4	5	6	7	8	9
Co-morbidities	Obese Insulin-dep. DM2 w/CKD COPD Nicotine AFib HTN HLD NSTEMI Pacemaker Long QT (1) h/o VTE (1) long-term AC ESRD on HD	HTN HLD	Kidney-heart transplant RA Protein-calorie malnutrition (1)	Obese HTN HLD HIV Alcoholic cirrhosis w/ascites (2)	Overweight (1) Prostate CA Alcoholic cirrhosis w/ascites Alcohol abuse (1) h/o pulm TB	Obese HTN HLD (2) Old MI (2) NSTEMI Long QT (1) ESRD on HD Protein-calorie malnutrition OSA	Insulin-dep. DM2 w/CKD AFib HTN HLD Long QT (2) Long-term AC CKD Kidney transplant Protein-calorie malnutrition (1)	Nicotine (2) AFib (2) HTN (2) Pacemaker (2) Long QT (2) h/o VTE (2) long-term AC (2)	AFib HTN HLD Old MI Pacemaker Long QT (2) Long-term AC CKD Breast cancer Protein-calorie malnutrition (2)

	Cancer - retroperitoneum								
Symptoms #1	SOB, diarrhea, weakness, low back pain, pneumonia, acute respiratory failure w/hypoxia, ARDS, altered mental status, metabolic encephalopathy, fluid overload,	SOB, cough, fever, chest pain, pneumonia, acute respiratory failure w/hypoxia, bradycardia	Cough, chest pain, pneumonia, acute kidney failure, tachycardia	Fever, tachycardia	Cough, fever, headache, chest pain, tachycardia, acute embolism and thrombosis (right femoral and unspecified)	Fever, diarrhea, pneumonia, acute respiratory failure with hypoxia, ARDS, ventilator dependence, acute kidney failure, encephalopathy, tachycardia, severe sepsis with shock, acute embolism	SOB, cough, headache, pneumonia, acute respiratory failure with hypoxia, ARDS, acute kidney failure, encephalopathy, fluid overload, sepsis with shock, acute embolism and thrombosis (unspecified)	NONE	Weakness, pneumonia, acute respiratory failure with hypoxia, acute kidney failure

	ventricular tachycardia					and thrombosis (right peroneal)			
Symptoms #2	acute respiratory failure w/hypoxia, ventricular tachycardia	NONE	SOB, diarrhea, respiratory failure w/hypoxia, acute kidney failure, tachycardia	SOB, fever, acute kidney failure, tachycardia acute embolism and thrombosis (unspecified)	Chest pain, tachycardia, acute embolism and thrombosis (unspecified)	Diarrhea, pneumonia, acute respiratory failure with hypoxia, ARDS, acute kidney failure, tachycardia, sepsis with shock, acute embolism and thrombosis (right peroneal)	Fever, pneumonia, acute respiratory failure with hypoxia, ventilator dependence, acute kidney failure, fluid overload, sepsis with shock	SOB, bradycardia	Diarrhea, weakness, pneumonia, acute respiratory failure with hypoxia, acute kidney failure

COVID dx	#1 yes, #2 yes	#1 yes, #2 NO	#1 NO, #2 yes	#1 yes, #2 yes	#1 yes, #2 NO	#1 yes, #2 yes	#1 yes, #2 yes	#1 NO, #2 yes	#1 yes, #2 yes
Treatment	#1: Hosp #2: Hosp	#1: Hosp, dexamethasone	#1: Hosp #2: Hosp	Hosp/dexa in between two positive tests and at time of second test	#1: Hosp (long)	#1: Hosp (long) #2: Hosp then dexa	#1: Hosp, tocilizumab	#1: Hosp (Mech Vent) #2: Hosp	#1: Hosp #2: Hosp, dexa
Notes		Pre-procedural exam (2)		Pre-procedural exam (1) I think this is all a single infection, hosp and toci somewhere in the middle of the two	Long hospitalization, many negative tests, one random second positive test right beside a negative; I think this is a single infection				2 COVID dx were only 18 days apart

*(# indicates if only coded for episode 1 or 2)

**immunocompromising condition

Abbreviations:

CA = cancer

MI = myocardial infarction

AC = anticoagulation

Immunocompromising conditions (N):

- ESRD on HD (4)
    - Prevalence of CKD in U.S. ~14%
    - ESRD incidence rate in 2013 was 363 per million/year, # of prevalence cases rises by ~21,000 cases per year
- HIV (3)
    - Prevalence in U.S. ~1.2 million[1]
- Transplant (3)
- Cirrhosis (5)
    - Prevalence in U.S. ~0.27%[2]
- Alcohol dependence or alcoholic cirrhosis (4)
- Cancer (4 solid, 2 liquid)
- Protein-calorie malnutrition (8)
- Rheumatologic / autoimmune (1)
- History of opportunistic lung infection (TB, Histoplasmosis, Blastomycosis; 3)

4.4 Discussion

In this study, we conducted a comprehensive characterization of COVID-19 patients with test re-positivity in a large EHR system across the US. We combined the clinical codes (ICD, LOINC) and a standardized concept set as our querying algorithm to best retrieve the patient information. We investigated the overall cumulative incidence rate, as well as the rate by age, gender, race, and BMI, of re-positivity in patients who are followed at least one day after their diagnosis date. From our result, we found that, male and aging lead to a higher risk of re-positivity. Although similar research on re-positivity is lacking, this finding is consistent with reports from other research, which showed these two factors are associated with an increased risk of severe COVID-19 outcomes [102, 103]. Against other research that obesity can increase the risk of disease severity and mortality, our finding shows that heavier weight provides some protective effects on COVID-19 re-positivity [104].

We further provide clinical and test characterization of 23 COVID-19 patients with suspected re-infection, defined as repeatedly positive SARS-CoV-2 PCR tests separated by consecutive negative tests and prolonged time. We observed a high prevalence of Black race, obesity, and multiple comorbidities known to increase risk of COVID-19

illness, including hypertension, diabetes and CKD. Moreover, 83% of the patients with repeated positivity were current smokers, which is linked to increased risk of severe COVID-19 [101]. It is possible that those known to be at particularly high risk for COVID-19 or those with persistent or recurrent symptoms may undergo frequent test, thereby increasing the likelihood of receiving some false positive or false negative results.

Immune compromising conditions including end-stage renal disease on dialysis, HIV, cirrhosis including alcohol-related, solid organ transplant, cancer, and protein-calorie malnutrition were more common in our study population affected by COVID-19 compared to the general population without COVID-19. Among the subset of patients in our study with immunocompromising conditions, more than two-thirds required hospitalization for the second positive PCR test after the interval negative PCR test. Reinfection may therefore raise clinical suspicion for an underlying immune defect, which may have also influenced duration to achieve viral clearance.

Recent studies focused on immune-compromised populations with COVID-19 have highlighted elevated risks of COVID-19 severity and morbidity, as well as frequent multimorbidity. High attributable 28-day mortality due to COVID-19 for patients in the ERA-EDTA Registry, including 3285 on dialysis and 1013 with a functional kidney

transplant, was substantial at 20% or 21-times higher and 19.9% or 92-times higher, respectively, compared to matched controls [105]. Similar trends of increased mortality and worsening liver function tests have been demonstrated in patients with cirrhosis and COVID-19 [106]. Moreover, alcohol-related liver disease and baseline hepatic dysfunction have been shown to be independent risk factors for death related to COVID-19 [107]. Patients with HIV in our sample had at least one other significant comorbidity, either alcoholic cirrhosis (PT4, PT14) or ESRD on HD (PT20). This aligns with findings of an observational prospective study in Madrid by Vizcarra et al. describing 51 HIV-positive patients with COVID-19 having a significantly higher prevalence of comorbidities and age-adjusted mortality [108].

Over 90% of patients in our study exhibited symptoms or clinical manifestations indicative of COVID-19 at the time of the second positive test, thereby fulfilling an important CDC criterion for the investigation of suspected SARS-CoV-2 reinfection, particularly among those with 45-89 days between positive PCR tests. Over three-fourths were hospitalized at the second positive test. Overall, 70% (12/17) of patients hospitalized at the first positive test were also hospitalized at the second test, suggesting that in most cases re-infection was not associated with less severe disease. This proportion is somewhat inconsistent with the finding of a recent review of 16 reported cases of re-infection confirmed by sequencing [109], in which the severity of the re-

infection episode was asymptomatic/mild in 75% of cases. Overall, in our study, 37% of those hospitalized had severe disease characterized by ICU admission. This is higher than previous estimates that 17-35% of hospitalized COVID-19 patients are treated in an ICU [98]. Acute kidney injury (AKI) has been reported to occur in approximately 9% of hospitalized COVID-19 patients and a higher proportion of those requiring ICU admission [98]. We observed AKI as a more common complication associated with COVID-19, including after both COVID-19 diagnoses in several individuals, but we were unable to determine whether these were independent or persistent events.

It is possible that hospitalized patients are more likely to undergo frequent test, due to more severe disease or to support discharge to a rehabilitation facility or nursing home. This frequent test can lead to alternating positive and negative tests, often on overlapping days. Given the high hospitalization rate in our study, repeated positive tests for some patients (e.g., PT5, PT7, PT12 and PT17) occurring during the time of an extended hospitalization with severe complications, including need for ICU admission and/or mechanical ventilation, may not represent true re-infections. Moreover, prolonged viral shedding, as has been observed in severe COVID-19 cases [110], cannot be ruled out. A recent analysis in the Emory Healthcare System indicated that, among 22,443 patients who had at least two tests, the median (IQR) duration between first and last positive test

was 19 days (12, 32), and a duration of 45 and 90 days represented the 88th and 97th percentile, respectively [109].

In the absence of genomic evaluations to definitively confirm reinfection [99, 109], finding two positive molecular tests separated by negative tests, prolonged time, and resolution of symptoms remains the best surrogate measure of possible re-infection. In 70 previously reported cases to date [99], with an average of 101 days between first and second positive test, viral genome sequences were shown to be distinct, strongly suggesting a re-infection rather than failure to clear an initial infection. Our identification and clinical characterization of 23 possible re-infections in a large dataset, with a median of 77 days between positive tests, provides additional data suggesting that re-infections may be common. Since most patients in the Optum dataset did not have repeated tests after their COVID-19 diagnosis, the true incidence rate of recurrent detectable SARS-CoV-2 cannot be estimated.

Our analysis was limited by lack of information on RT-PCR platforms (with varying sensitivities) or semi-quantitative RT-PCR cycle threshold (Ct) values. The patients in our study nevertheless fulfilled CDC criteria for cases >90 days apart or 45-89 days apart based on positive RT-PCR, and based on our study definition, cases were classified as re-infection rather than relapse based on interval negative RT-PCR. We were not able to

confirm if COVID-19 was the primary diagnosis prompting hospitalization, if patients were incidentally found to test positive for SARS-CoV-2 upon admission for an unrelated illness, or later became symptomatic during the hospitalization course. Diagnoses may be more likely to be incidental if associated with ICD-10 codes for pre-procedural exam (e.g., elective surgery); however, a pre-procedural exam at the time of the second positive test was noted for only one patient in our study (PT2). It is noteworthy that among the 23 patients with confirmed RT-PCR re-positivity for SARS-CoV-2, a minority were not assigned an ICD-10 code for COVID-19 in the EHR. This did not appear to be associated with severity of disease presentation. Finally, repeatedly positive tests do not necessarily mean a re-infection, and persistent infection or relapse cannot be ruled out, particularly if signs and symptoms observed at the second positive test are similar to those seen in individuals with post-acute sequelae of COVID-19 (or "long COVID").

Despite these limitations, our study provides a comprehensive characterization of demographic, clinical and SARS-CoV-2 test data for patients with repeatedly positive SARS-CoV-2 tests in a large EHR database across the US, which could help prioritize suspected cases of reinfection for investigation in the absence of sequencing data and for continued surveillance for potential long-term health consequences of SARS-CoV-2 infection. Further investigation into risk of reinfection by type and degree of

immunosuppressive condition, medications, and disease chronicity will be valuable for future goals of prevention, mitigation of risk factors, and reducing severity of illness.

4.5 Conclusion

Our study is an implementation of medical informatics to real-world study to improve the research efficiency. This study demonstrated a high prevalence of immune compromise, comorbidities, obesity and smoking among patients with repeatedly positive SARS-CoV-2 tests, which are comparable with findings from other studies. Despite limitations, including lack of semi-quantitative estimates of viral load, these data may help prioritize suspected cases of reinfection for investigation and continued surveillance.

Chapter 5: Conclusion

5.1 Summary of key findings

This is a study to explore the roles and implementation of medical informatics technology, specifically NLP and ontology methods, in standardizing the emerging lab tests during a pandemic, thus to facilitate rapid responses to the pandemic. The ultimate goal of this study is to construct an informatics framework for rapid standardization of lab tests during a pandemic, thus to better prepare for future public health threats. We first developed an information model for lab tests approved during the COVID-19 pandemic and built an NER tool that can automatically extract lab test information specified in the information model from the FDA EUAs documents, thus creating a catalog of approved lab tests with detailed information. To foster standardization of lab testing data in EHRs, we further developed the COVID-19 TestNorm, a tool to normalize various COVID-19 lab testing names used by different healthcare facilities into the standard LOINC codes. Finally, we conducted a clinical research on COVID-19 re-positivity to demonstrate the utility of standardized lab test information in a pivotal clinical research of COVID-19. The main findings of each chapter are summarized below.

In chapter 2, we took COVID-19 lab tests as a use case to design an information model for lab tests developed during a pandemic, as well as to develop NER methods to automatically extract standardized information from the FDA EUAs documents. We collected 378 COVID-19 lab test EUAs from FDA and annotated them according to the types of entities specified in the COVID-19 lab test information model. We then developed, evaluated and compared three NER models (the baseline CRF model, the BERT model and the Bi-LSTM-CRF model) on the corpus. Our results indicate that both the BERT model and the Bi-LSTM-CRF model achieve remarkable performance in the NER task, which demonstrates the utility of NLP and ontology technologies for extracting standardized information from EUAs in pandemic emergencies.

In chapter 3, we collected lab test results from eight healthcare systems across the country and developed a simple but effective normalization system for mapping COVID-19 lab tests results to LOINC codes, in order to facilitate a rapid research response to the pandemic. The overall accuracy of COVID-19 TestNorm on the development set was 98.9%, and on the independent test set was 97.4%. The tool is publicly available with source code. For ease of use, a web application has been developed to enable end users to use the online form and easily map their local COVID-19 lab test names to standardized LOINC codes, thereby enhancing the efficiency of multi-center data aggregation and

global knowledge sharing, furnishing a fundamental step towards test data interoperability for research related to COVID-19.

In chapter 4, we conducted a comprehensive characterization of COVID-19 patients with test re-positivity in a large EHR system across the US. We combined the clinical codes (ICD, LOINC) and a standardized concept set as our query algorithm to best retrieve the patient information from the dataset. We investigated the overall cumulative incidence, and the rate of re-positivity by age, gender, race and BMI, in patients who were followed up at least one day after the date of diagnosis. From our result, we find that being male and aging lead to a higher risk of re-positivity. For all the lack of similar research on re-positivity, this finding is consistent with reports from other research, which indicate that both factors are associated with an increased risk of severe COVID-19 outcomes. Contrary to other research suggesting that obesity can raise the risk of disease severity and mortality, our finding demonstrates that heavier weight offers some protection against COVID-19 re-positivity. We provide the clinical and detection traits of 23 patients with COVID-19 suspected of reinfection, which was defined as repeated positive SARS-CoV-2 PCR detection, continuous negative detection and long-term interlude. We observed that Black race, obesity, and a variety of illnesses, including hypertension, diabetes and CKD, would lead to high incidence rates of complications of COVID-19. Moreover, 83% of the patients with repositive detection were currently smokers, which

was associated with an increased risk of severe COVID-19 illness. The results are comparable to those of other traditional clinical researches. This study demonstrates the feasibility of using a standardized vocabulary in clinical researches to enhance the ease and efficiency of data preparation and analysis, and ultimately facilitate rapid response to the pandemic outbreak.

5.2 Innovations and contributions

5.2.1 Innovations

To the best of our knowledge, this is the first study to construct an information model to represent lab tests developed during a pandemic. We identified a number of challenges, and applied a range of innovative informatics approaches to address them, including:

a. A new information model was specifically designed to represent COVID-19 lab tests, with general applications to future pandemics in mind. In spite of the existing COVID-19 related ontologies, they focus more on disease development, etiology or the structure of clinical documentations. Our model concentrates on the standardization of lab test information, which furnishes a set of attributes to represent specific information about a COVID-19 lab test.

b. A high-performance, machine-learning based information extraction model to parse the FDA EUAs. This is the first attempt to parse the EUAs since these documents are intended for public health emergencies. During the COVID-19 outbreak, all the lab tests, as well as vaccines, should first be authorized through EUAs before they can be marketed in the U.S. This process is dramatically different from the regular approval process of a lab test. Hence, our high-performance NER model is the first and only one that can efficiently and accurately extract information from EUAs.

c. An easy-to-use automated tool that normalizes local COVID-19 test names into standard LOINC codes. This tool is the first publicly available COVID-19 lab test normalization tool, which has been adopted by a number of initiatives.

5.2.2 Contributions

This work contributes to both the biomedical informatics and clinical practice/research fields in the following aspects.

a. Our study established a unique lab test ontology for COVID-19, which has great potential for scalability and generalizability in future pandemic outbreaks. This

ontology affords a validated lab test for disambiguation in clinical practice and COVID-19 related research.

b. Our study developed a high-performance NER model to extract information from EUAs. From the research on the lab test development and authorization process during COVID-19, it can also be expected that future pandemics will lead to a similar lab test authorization process, and therefore our machine-learning based NER model is of significant extendibility and generalizability for future responses to potential public health threats.

c. Our publicly available, easy-to-use COVID-19 lab test normalization tool is friendly to local small health care facilities, allowing them to map their lab test names to the standard vocabulary. This offers interoperability of test data for research related to COVID-19.

d. We conducted a comprehensive characterization of COVID-19 patients with test re-positivity in a large-scale real-world dataset across the U.S. In spite of the studies exploring the phenomenon of re-infection, most confirmed cases are sparse, and the study cohorts are small. In our study, we used a large-scale real-

world database to build our study cohort. We employed the hybrid searching algorithm that combined LOINC, ICD-10 codes with the standardized set of concepts derived from the COVID-19 lab test ontology, in order to improve the information coverage of COVID-19 patients. This study demonstrates the significant potential of applying standardized vocabulary to real world datasets to enhance the efficiency and data quality of real-world study.

5.3 Limitations and future work

As a preliminary study exploring how NLP and ontology technologies can be applied to standardize lab testing data in a pandemic, this study has several limitations. First, COVID-19 is the only use case for developing the information model and NER tool with the final goal of representing lab tests in general pandemics. Although we take into consideration other previous pandemics during the information model development, the lack of emergency documents for those previous pandemic lab tests hinders us from a holistic picture of how the lab tests emerged during those periods. Our future work is to collaborate with domain experts to further test and optimize the ontology, entitling it higher potential to be generalized to future pandemics. Second, the sample size and data heterogeneity are still limited, even though we have collected data from eight large healthcare systems across the United States. For example, all codes in our dataset are

about molecular and antibody detection. With new test results available on the market, the LOINC code sets for COVID-19 continues to evolve, i.e., weekly updates from Regenstrief and ongoing updates from the CDC, which maintains a file with recommended LOINC mappings relationships for currently FDA-approved test kits (https://www.cdc.gov/csels/dls/sars-cov-2-livd-codes.html). Therefore, it is critical for us to constantly update our tool with new code sets and coding rules. When a large number of diverse samples have been accumulated, we will also investigate more sophisticated machine learning approaches to accomplish this task. Third, for the real-world re-positivity characterization, our analysis was limited due to the lack of information on RT-PCR platforms (with varying sensitivities) or semi-quantitative RT-PCR cycle threshold (Ct) values. There is also some other clinical information not included in the real-world dataset. By employing the hybrid searching strategy, we extended the information coverage of the selected patients, but our study is still limited by the natural deficiencies of a real-world study. Therefore, future close collaborations with domain experts, such as clinicians and epidemiologists, will contribute to the optimization of our study and enhance its validity.

5.4 Conclusion

In this work, taking the COVID-19 as a use case, I developed informatics approaches for standardizing lab tests, and demonstrate the application of standardized lab test

information in clinical research. The results of my study indicate the great potential of medical informatics technologies in facilitating rapid response to both current and future pandemics.

References

1. CDC. (2020, February 11). Healthcare Workers. *Centers for Disease Control and Prevention*. Retrieved June 27, 2021, from https://www.cdc.gov/coronavirus/2019-ncov/hcp/testing-overview.html
2. Rosenthal, P. J. (2020). The Importance of Diagnostic Testing during a Viral Pandemic: Early Lessons from Novel Coronavirus Disease (COVID-19). *The American Journal of Tropical Medicine and Hygiene*, *102*(5), 915–916. https://doi.org/10.4269/ajtmh.20-0216
3. Morens, D. M., & Fauci, A. S. (2020). Emerging Pandemic Diseases: How We Got to COVID-19. *Cell*, *182*(5), 1077–1092. https://doi.org/10.1016/j.cell.2020.08.021
4. Sahajpal, N. S., Njau, A., Mondal, A. K., Ananth, S., Chaubey, A., Rojiani, A., & Kolhe, R. (n.d.). Role of clinical laboratories in response to the COVID-19 pandemic. *Future Medicinal Chemistry*, 10.4155/fmc-2020–0129. https://doi.org/10.4155/fmc-2020-0129
5. Commissioner, O. of the. (2021). Emergency Use Authorization. *FDA*. Retrieved from https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/emergency-use-authorization
6. The trouble with (non) standardized tests. (2020, June 25). *Healthcare Purchasing News*. Retrieved June 27, 2021, from

https://www.hpnonline.com/sourcing-logistics/article/21142737/the-trouble-with-non-standardized-tests

7. Zitek, T. (2020). The Appropriate Use of Testing for COVID-19. *The Western Journal of Emergency Medicine*, *21*(3), 470–472. https://doi.org/10.5811/westjem.2020.4.47370
8. COVID-19 Pandemic Response, Laboratory Data Reporting: CARES Act Section 18115. (n.d.), 7.
9. Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & Metabolic Syndrome*, *14*(4), 337–339. https://doi.org/10.1016/j.dsx.2020.04.012
10. Hu, Z., Ge, Q., Li, S., Jin, L., & Xiong, M. (2020). Artificial Intelligence Forecasting of Covid-19 in China. *arXiv:2002.07112 [q-bio]*. Retrieved from http://arxiv.org/abs/2002.07112
11. Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases | Radiology. (n.d.). Retrieved June 28, 2021, from https://pubs.rsna.org/doi/10.1148/radiol.2020200642
12. Stebbing, J., Phelan, A., Griffin, I., Tucker, C., Oechsle, O., Smith, D., & Richardson, P. (2020). COVID-19: combining antiviral and anti-inflammatory treatments. *The Lancet Infectious Diseases*, *20*(4), 400–402. https://doi.org/10.1016/S1473-3099(20)30132-8

13. Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P. D., Zhang, H., Ji, W., … Siegel, E. (2020). Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis. *arXiv:2003.05037 [cs, eess]*. Retrieved from http://arxiv.org/abs/2003.05037
14. Ting, D. S. W., Carin, L., Dzau, V., & Wong, T. Y. (2020). Digital technology and COVID-19. *Nature Medicine*, *26*(4), 459–461. https://doi.org/10.1038/s41591-020-0824-5
15. About LOINC. (n.d.). *LOINC*. Retrieved from https://loinc.org/about/
16. SNOMED CT. (n.d.). Product, Program, and Project Descriptions, U.S. National Library of Medicine. Retrieved June 28, 2021, from https://www.nlm.nih.gov/healthit/snomedct/index.html
17. Lafferty, J., McCallum, A., & Pereira, F. C. N. (n.d.). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, 10.
18. Panchendrarajan, R., & Amaresan, A. (2018). Bidirectional LSTM-CRF for Named Entity Recognition. *Information and Computation*, 10.
19. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. Retrieved from http://arxiv.org/abs/1810.04805

20. *ICD-10: International statistical classification of diseases and related health problems.* (2011). Geneva: World Health Organization.
21. CDC. (2020, February 11). Coronavirus Disease 2019 (COVID-19). *Centers for Disease Control and Prevention*. Retrieved June 28, 2021, from https://www.cdc.gov/coronavirus/2019-ncov/science/about-epidemiology/identifying-source-outbreak.html
22. Global HIV & AIDS statistics — Fact sheet. (n.d.). Retrieved June 28, 2021, from https://www.unaids.org/en/resources/fact-sheet
23. Fournier, J.-M., & Quilici, M.-L. (2007). [Cholera]. *Presse Medicale (Paris, France: 1983)*, *36*(4 Pt 2), 727–739. https://doi.org/10.1016/j.lpm.2006.11.029
24. Zietz, B. P., & Dunkelberg, H. (2004). The history of the plague and the research on the causative agent Yersinia pestis. *International Journal of Hygiene and Environmental Health*, *207*(2), 165–178. https://doi.org/10.1078/1438-4639-00259
25. History of 1918 Flu Pandemic | Pandemic Influenza (Flu) | CDC. (2019, January 22). Retrieved June 28, 2021, from https://www.cdc.gov/flu/pandemic-resources/1918-commemoration/1918-pandemic-history.htm
26. Jester, B. J., Uyeki, T. M., & Jernigan, D. B. (2020). Fifty Years of Influenza A(H3N2) Following the Pandemic of 1968. *American Journal of Public Health*, *110*(5), 669–676. https://doi.org/10.2105/AJPH.2019.305557

27. Zhong, N., Zheng, B., Li, Y., Poon, L., Xie, Z., Chan, K., … Guan, Y. (2003). Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *Lancet (London, England)*, *362*(9393), 1353–1358. https://doi.org/10.1016/S0140-6736(03)14630-2
28. Memish, Z. A., Perlman, S., Van Kerkhove, M. D., & Zumla, A. (2020). Middle East respiratory syndrome. *Lancet (London, England)*, *395*(10229), 1063–1077. https://doi.org/10.1016/S0140-6736(19)33221-0
29. Hu, T., Liu, Y., Zhao, M., Zhuang, Q., Xu, L., & He, Q. (2020). A comparison of COVID-19, SARS and MERS. *PeerJ*, *8*, e9725. https://doi.org/10.7717/peerj.9725
30. Pride, D. (n.d.). There are many COVID-19 tests in the US – how are they being regulated? *The Conversation*. Retrieved June 28, 2021, from http://theconversation.com/there-are-many-covid-19-tests-in-the-us-how-are-they-being-regulated-134783
31. Ravi, N., Cortade, D. L., Ng, E., & Wang, S. X. (2020). Diagnostics for SARS-CoV-2 detection: A comprehensive review of the FDA-EUA COVID-19 testing landscape. *Biosensors & Bioelectronics*, *165*, 112454. https://doi.org/10.1016/j.bios.2020.112454
32. Goudouris, E. S. (2021). Laboratory diagnosis of COVID-19. *Jornal De Pediatria*, *97*(1), 7–12. https://doi.org/10.1016/j.jped.2020.08.001

33. St Hilaire, B. G., Durand, N. C., Mitra, N., Pulido, S. G., Mahajan, R., Blackburn, A., … Dudchenko, O. (2020). A rapid, low cost, and highly sensitive SARS-CoV-2 diagnostic based on whole genome sequencing. *bioRxiv*.

34. Wolters, F., van de Bovenkamp, J., van den Bosch, B., van den Brink, S., Broeders, M., Chung, N. H., … Meijer, A. (2020). Multi-center evaluation of cepheid xpert® xpress SARS-CoV-2 point-of-care test during the SARS-CoV-2 pandemic. *Journal of Clinical Virology*, *128*, 104426. https://doi.org/10.1016/j.jcv.2020.104426

35. Moran, A., Beavis, K. G., Matushek, S. M., Ciaglia, C., Francois, N., Tesic, V., & Love, N. (n.d.). Detection of SARS-CoV-2 by Use of the Cepheid Xpert Xpress SARS-CoV-2 and Roche cobas SARS-CoV-2 Assays. *Journal of Clinical Microbiology*, *58*(8), e00772-20. https://doi.org/10.1128/JCM.00772-20

36. Bordi, L., Piralla, A., Lalle, E., Giardina, F., Colavita, F., Tallarita, M., … Capobianchi, M. R. (2020). Rapid and sensitive detection of SARS-CoV-2 RNA using the Simplexa™ COVID-19 direct assay. *Journal of Clinical Virology*, *128*, 104416. https://doi.org/10.1016/j.jcv.2020.104416

37. The Basics: RNA Isolation - US. (n.d.). Retrieved June 28, 2021, from //www.thermofisher.com/us/en/home/references/ambion-tech-support/rna-isolation/general-articles/the-basics-rna-isolation.html

38. PerkinElmer, I. (2020). *PerkinElmer® New Coronavirus Nucleic Acid Detection Kit V 2.0*.

39. Yang, Y., Yang, M., Shen, C., Wang, F., Yuan, J., Li, J., … Liu, Y. (2020). Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections. *medRxiv*, 2020.02.11.20021493. https://doi.org/10.1101/2020.02.11.20021493
40. Scientific, T. F. (2020). *TaqPath TM COVID-19 Combo Kit*.
41. Dutta, N. K., Mazumdar, K., & Gordy, J. T. (2020). The Nucleocapsid Protein of SARS–CoV-2: a Target for Vaccine Development. *Journal of Virology*, *94*(13), e00647-20. https://doi.org/10.1128/JVI.00647-20
42. Koczula, K. M., & Gallotta, A. (2016). Lateral flow assays. *Essays in Biochemistry*, *60*(1), 111–120. https://doi.org/10.1042/EBC20150012
43. British Society for immunology. (n.d.). Enzyme-linked immunosorbent assay (ELISA). *Immunology*.
44. Complete Solutions for IVD Chemiluminescent Immunoassay (CLIA/CLEIA) Development. (n.d.). Retrieved June 28, 2021, from https://www.sigmaaldrich.com/US/en/technical-documents/technical-article/clinical-testing-and-diagnostics-manufacturing/ivd-manufacturing/clia-chemiluminescent-immunoassay-development
45. Monobind.com: CLIA Advantages. (n.d.). Retrieved June 28, 2021, from https://www.monobind.com/Additional-Info/Assays-CLIA-Advantages

46. Clinical Laboratory Improvement Amendments (CLIA) | CMS. (n.d.). Retrieved June 28, 2021, from https://www.cms.gov/Regulations-and-Guidance/Legislation/CLIA
47. Most Coronavirus Tests Cost About $100. Why Did One Cost $2,315? - The New York Times. (n.d.). Retrieved June 28, 2021, from https://www.nytimes.com/2020/06/16/upshot/coronavirus-test-cost-varies-widely.html
48. Barton, C. M., Alberti, M., Ames, D., Atkinson, J.-A., Bales, J., Burke, E., … Tucker, G. (2020). Call for transparency of COVID-19 models. *Science (New York, N.Y.)*, *368*(6490), 482–483. https://doi.org/10.1126/science.abb8637
49. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., … Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, *25*(11), 1251. https://doi.org/10.1038/nbt1346
50. Eilbeck, K., Jacobs, J., McGarvey, S., Vinion, C., & Staes, C. (2013). Exploring the use of ontologies and automated reasoning to manage selection of reportable condition lab tests from LOINC. *CEUR Workshop Proceedings*, *1060*, 12–15.
51. Council of State and Territorial Epidemiologists. (n.d.). Retrieved June 28, 2021, from https://www.cste.org/group/RCKMS
52. HermiT Reasoner: Home. (n.d.). Retrieved June 28, 2021, from http://www.hermit-reasoner.com/

53. Campbell, W. S., Karlsson, D., Vreeman, D. J., Lazenby, A. J., Talmon, G. A., & Campbell, J. R. (2018). A computable pathology report for precision medicine: extending an observables ontology unifying SNOMED CT and LOINC. *Journal of the American Medical Informatics Association*, *25*(3), 259–266. https://doi.org/10.1093/jamia/ocx097

54. He, Y., Yu, H., Ong, E., Wang, Y., Liu, Y., Huffman, A., … Smith, B. (2020). CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Scientific Data*, *7*(1), 181. https://doi.org/10.1038/s41597-020-0523-6

55. Chemical Entities of Biological Interest (ChEBI). (n.d.). Retrieved June 28, 2021, from https://www.ebi.ac.uk/chebi/

56. Human Phenotype Ontology. (n.d.). Retrieved June 28, 2021, from https://hpo.jax.org/app/

57. Disease Ontology - Institute for Genome Sciences @ University of Maryland. (n.d.). Retrieved June 28, 2021, from https://disease-ontology.org/

58. Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, *40*(Database issue), D136–D143. https://doi.org/10.1093/nar/gkr1178

59. Sarawagi, S. (2008). *Information Extraction*. Now Publishers Inc.

60. Small, S. G., & Medsker, L. (2014). Review of information extraction technologies and applications. *Neural Computing and Applications*, *25*(3), 533–548. https://doi.org/10.1007/s00521-013-1516-6

61. Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM*, *39*(1), 80–91. https://doi.org/10.1145/234173.234209

62. Nadeau, D., & Sekine, S. (n.d.). A survey of named entity recognition and classification, 20.

63. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task* (pp. 28–34). Presented at the Conference on Computational Natural Language Learning, Date: 2011/06/23 - 2011/06/24, Location: Portland, Oregon, Association for Computational Linguistics. Retrieved from https://lirias.kuleuven.be/1821222

64. Bach, N., & Badaskar, S. (n.d.). A Review of Relation Extraction, 15.

65. Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., … Liu, H. (2018). Clinical Information Extraction Applications: A Literature Review. *Journal of biomedical informatics*, *77*, 34–49. https://doi.org/10.1016/j.jbi.2017.11.011

66. Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction

System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association: JAMIA*, *17*(5), 507–513. https://doi.org/10.1136/jamia.2009.001560

67. Aronson, A. R., & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, *17*(3), 229–236. https://doi.org/10.1136/jamia.2009.002733
68. Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association: JAMIA*, *1*(2), 161–174. https://doi.org/10.1136/jamia.1994.95236146
69. Unified Medical Language System (UMLS). (n.d.). List of Links, U.S. National Library of Medicine. Retrieved June 28, 2021, from https://www.nlm.nih.gov/research/umls/index.html
70. RadLex radiology lexicon. (n.d.). Retrieved June 28, 2021, from https://www.rsna.org/practice-tools/data-tools-and-standards/radlex-radiology-lexicon
71. NLP's ImageNet moment has arrived. (n.d.). Retrieved June 28, 2021, from https://ruder.io/nlp-imagenet/
72. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

73. Patrick, E. A., & Fischer, F. P. (1970). A generalized k-nearest neighbor rule. *Information and Control*, *16*(2), 128–152. https://doi.org/10.1016/S0019-9958(70)90081-1

74. Valueva, M. V., Nagornov, N. N., Lyakhov, P. A., Valuev, G. V., & Chervyakov, N. I. (2020). Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation*, *177*, 232–243. https://doi.org/10.1016/j.matcom.2020.04.031

75. Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability | Wiley. (n.d.). *Wiley.com*. Retrieved June 28, 2021, from https://www.wiley.com/en-us/Recurrent+Neural+Networks+for+Prediction%3A+Learning+Algorithms%2C+Architectures+and+Stability-p-9780471495178

76. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural computation*, *9*, 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

77. Si, Y., Wang, J., Xu, H., & Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association: JAMIA*, *26*(11), 1297–1304. https://doi.org/10.1093/jamia/ocz096

78. Miller, G. (n.d.). Standardization of Laboratory Tests - How to do it, 59.

79. Vesper, H. W., & Thienpont, L. M. (2009). Traceability in laboratory medicine. *Clinical Chemistry*, *55*(6), 1067–1075. https://doi.org/10.1373/clinchem.2008.107052

80. ISO - ISO 17511:2003 - In vitro diagnostic medical devices — Measurement of quantities in biological samples — Metrological traceability of values assigned to calibrators and control materials. (n.d.). Retrieved June 28, 2021, from https://www.iso.org/standard/30716.html

81. About Health Level Seven International | HL7 International. (n.d.). Retrieved June 28, 2021, from http://www.hl7.org/about/index.cfm?ref=nav

82. Haendel, M. A., Chute, C. G., Bennett, T. D., Eichmann, D. A., Guinney, J., Kibbe, W. A., … N3C Consortium. (2021). The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *Journal of the American Medical Informatics Association: JAMIA*, *28*(3), 427–443. https://doi.org/10.1093/jamia/ocaa196

83. Brat, G. A., Weber, G. M., Gehlenborg, N., Avillach, P., Palmer, N. P., Chiovato, L., … Kohane, I. S. (2020). International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ digital medicine*, *3*, 109. https://doi.org/10.1038/s41746-020-00308-0

84. Guidance for mapping to SARS-CoV-2 LOINC terms. (n.d.). *LOINC*. Retrieved from https://loinc.org/sars-coronavirus-2/

85. Smallpox | CDC. (2019, February 19). Retrieved June 28, 2021, from https://www.cdc.gov/smallpox/index.html

86. Morabia, A. (2020). Pandemics and methodological developments in epidemiology history. *Journal of Clinical Epidemiology*, *125*, 164–169. https://doi.org/10.1016/j.jclinepi.2020.06.008

87. Hassan, E. M., & Mahmoud, H. N. (2021). Impact of multiple waves of COVID-19 on healthcare networks in the United States. *PLOS ONE*, *16*(3), e0247463. https://doi.org/10.1371/journal.pone.0247463

88. Amith, M., Roberts, K., & Tao, C. (2019). Conceiving an application ontology to model patient human papillomavirus vaccine counseling for dialogue management. *BMC Bioinformatics*, *20*(21), 706. https://doi.org/10.1186/s12859-019-3193-7

89. Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. (2018). CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, *25*(3), 331–336. https://doi.org/10.1093/jamia/ocx132

90. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv:1508.01991 [cs]*. Retrieved from http://arxiv.org/abs/1508.01991

91. softmax function. (n.d.). Retrieved June 28, 2021, from https://www.deeplearningbook.org/contents/mlp.html

92. lemonhu. (2021). *lemonhu/NER-BERT-pytorch*. Python. Retrieved from https://github.com/lemonhu/NER-BERT-pytorch

93. Huang, L.-C., Soysal, E., Zheng, W. J., Zhao, Z., Xu, H., & Sun, J. (2015). A weighted and integrated drug-target interactome: drug repurposing for schizophrenia as a use case. *BMC Systems Biology*, *9*(4), S2. https://doi.org/10.1186/1752-0509-9-S4-S2

94. Ezen-Can, A. (2020). A Comparison of LSTM and BERT for Small Corpus. *arXiv:2009.05451 [cs]*. Retrieved from http://arxiv.org/abs/2009.05451

95. OHDSI – Observational Health Data Sciences and Informatics. (n.d.). Retrieved from https://ohdsi.org/

96. *UTHealth-CCB/covid19_testnorm*. (2020). Python, UTHealth-CCB. Retrieved from https://github.com/UTHealth-CCB/covid19_testnorm

97. Wajnberg, A., Amanat, F., Firpo, A., Altman, D. R., Bailey, M. J., Mansour, M., … Cordon-Cardo, C. (2020). Robust neutralizing antibodies to SARS-CoV-2 infection persist for months. *Science*, *370*(6521), 1227–1230. https://doi.org/10.1126/science.abd7728

98. Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J., & Prescott, H. C. (2020). Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA*, *324*(8), 782–793. https://doi.org/10.1001/jama.2020.12839

99. COVID-19 reinfection tracker - BNO News. (n.d.). Retrieved June 28, 2021, from https://bnonews.com/index.php/2020/08/covid-19-reinfection-tracker/

100. CDC. (2020, May 29). Coronavirus Disease 2019 (COVID-19) in the U.S. *Centers for Disease Control and Prevention*. Retrieved June 2, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

101. COVID-19 Data Available to Researchers | Research Data Portal. (n.d.). Retrieved June 28, 2021, from https://dataportal.uta.edu/node/27

102. Peckham, H., de Gruijter, N. M., Raine, C., Radziszewska, A., Ciurtin, C., Wedderburn, L. R., … Deakin, C. T. (2020). Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. *Nature Communications*, *11*(1), 6317. https://doi.org/10.1038/s41467-020-19741-6

103. Gallo Marin, B., Aghagoli, G., Lavine, K., Yang, L., Siff, E. J., Chiang, S. S., … Michelow, I. C. (2021). Predictors of COVID-19 severity: A literature review. *Reviews in Medical Virology*, *31*(1), 1–10. https://doi.org/10.1002/rmv.2146

104. Sanchis-Gomar, F., Lavie, C. J., Mehra, M. R., Henry, B. M., & Lippi, G. (2020). Obesity and Outcomes in COVID-19: When an Epidemic and Pandemic Collide. *Mayo Clinic Proceedings*, *95*(7), 1445–1453. https://doi.org/10.1016/j.mayocp.2020.05.006

105. Jager, K. J., Kramer, A., Chesnaye, N. C., Couchoud, C., Sánchez-Álvarez, J. E., Garneata, L., … Massy, Z. A. (2020). Results from the ERA-EDTA Registry indicate a high mortality due to COVID-19 in dialysis patients and kidney transplant

recipients across Europe. *Kidney International*, *98*(6), 1540–1548. https://doi.org/10.1016/j.kint.2020.09.006

106. Iavarone, M., D'Ambrosio, R., Soria, A., Triolo, M., Pugliese, N., Del Poggio, P., … Lampertico, P. (2020). High rates of 30-day mortality in patients with cirrhosis and COVID-19. *Journal of Hepatology*, *73*(5), 1063–1071. https://doi.org/10.1016/j.jhep.2020.06.001

107. Marjot, T., Moon, A. M., Cook, J. A., Abd-Elsalam, S., Aloman, C., Armstrong, M. J., … Webb, G. J. (2021). Outcomes following SARS-CoV-2 infection in patients with chronic liver disease: An international registry study. *Journal of Hepatology*, *74*(3), 567–577. https://doi.org/10.1016/j.jhep.2020.09.024

108. Vizcarra, P., Pérez-Elías, M. J., Quereda, C., Moreno, A., Vivancos, M. J., Dronda, F., … Vizcarra, P. (2020). Description of COVID-19 in HIV-infected individuals: a single-centre, prospective cohort. *The Lancet HIV*, *7*(8), e554–e564. https://doi.org/10.1016/S2352-3018(20)30164-8

109. The Importance and Challenges of Identifying SARS-CoV-2 Reinfections | Journal of Clinical Microbiology. (n.d.). Retrieved June 28, 2021, from https://journals.asm.org/doi/full/10.1128/JCM.02769-20

110. Cevik, M., Tate, M., Lloyd, O., Maraolo, A. E., Schafers, J., & Ho, A. (2021). SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral

shedding, and infectiousness: a systematic review and meta-analysis. *The Lancet Microbe*, *2*(1), e13–e22. https://doi.org/10.1016/S2666-5247(20)30172-5