Mining Concepts for a COVID Interface Terminology for Annotation of EHRs

Vipina K. Keloth
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ USA
vk396@njit.edu

Shuxin Zhou
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ USA
sz23@njit.edu

Luke Lindemann
Yale Center for Medical Informatics
Yale University
New Haven, CT USA
luke.lindemann@gmail.com

Gai Elhanan
Renown Institute for Health Innovation
Desert Research Institute
Reno, NV USA
gelhanan@gmail.com

Andrew J. Einstein
Dept. of Medicine, Cardiology Division, and Dept. of Radiology
Columbia University Irving Medical Center
New York, NY USA
ae2214@cumc.columbia.edu

James Geller
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ USA
james.geller@njit.edu

Yehoshua Perl
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ USA
yehoshua.perl@njit.edu

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9377981

***Abstract*— The COVID-19 pandemic has overwhelmed the healthcare services of many countries with increased number of patients and also with a deluge of medical data. Furthermore, the emergence and global spread of new infectious diseases are highly likely to continue in the future. Incomplete data about presentations, signs, and symptoms of COVID-19 has had adverse effects on healthcare delivery. The EHRs of US hospitals have ingested huge volumes of relevant, up-to-date data about patients, but the lack of a proper system to annotate this data has greatly reduced its usefulness. We propose to design a COVID *interface terminology* for the annotation of EHR notes of COVID-19 patients. The initial version of this interface terminology was created by integrating COVID concepts from existing ontologies. Further enrichment of the interface terminology is performed by mining high granularity concepts from EHRs, because such concepts are usually not present in the existing reference terminologies. We use the techniques of *concatenation* and *anchoring* iteratively to extract high granularity phrases from the clinical text. In addition to increasing the conceptual base of the COVID interface terminology, this will also help in generating training data for large scale *concept mining* using machine learning techniques. Having the annotated clinical notes of COVID-19 patients available will help in speeding up research in this field.**

Keywords—COVID-19 ontologies, interface terminology, COVID-19 patient EHRs, concept mining, EHR annotation

I. INTRODUCTION

COVID-19 has turned into the greatest healthcare challenge since the Spanish flu pandemic, causing millions of infections and over one million deaths. Meanwhile, the Electronic Health Records (EHRs) in hospitals are ingesting a deluge of COVID-19 cases and morbidity information. COVID-19 uncovered weaknesses in US health information management practices that hamper research on the disease. At the early stages of this pandemic, doctors have been describing signs and symptoms in various organ systems, e.g., "COVID toes" and Multisystem Inflammatory Syndrome in Children (MIS-C). However, most of these terms are not coded and are only recorded as free text, inhibiting interoperability, and the use of EHR notes for research on the disease. How can we support research on "COVID toes" and other related COVID-19 rashes (for example), if we cannot code such findings in the EHR to make them easily discoverable, and doctors and clinical software are forced to search for them as free text instead of as concepts?

Clinical text in the form of consultation notes, nursing records, progress notes, etc. are reflections of changes in patient conditions and can provide relevant information to deliver better healthcare outcomes. However, the information in clinical notes may not exist as coded data in the EHR's problem list, or in encounter, admissions, or discharge diagnoses. This missing information could be extremely valuable, especially at the time of an emerging disease such as COVID-19 with a marked lack of information on its presentation, signs, symptoms, clinical progressions, and outcomes. A solution would be to annotate the clinical notes with concepts from an interface terminology to facilitate the extraction of such data in near real-time.

In reality, there are currently no satisfactory tools that enable clinical note annotation, partly because physicians record clinical notes with medical phrases that correspond to chunks in Cognitive Psychology [1]. These cognitive chunks are high granularity phrases many of which are not contained in standard medical reference terminologies used for annotation. Medical acronyms, found extensively in EHRs, are often indicative of such chunks. Some other examples from COVID-19 radiology reports are *extensive areas of crazy-paving patterns*, *bilateral parenchymal thickening*, *thickening of the interlobular septa*, etc. The widely used medical reference terminologies do not contain these chunks, but instead contain specific concepts such as *bilateral*, *thickness*, *interlobular*, *septa*, etc. Hence annotating with general purpose annotation tools using reference terminologies or the UMLS Metathesaurus [2] captures only fractional information.

In this paper, we propose to create a COVID Interface Terminology (CIT) to annotate EHRs of COVID-19 patients. An interface terminology [3] is different from a reference terminology. The former is designed to maximize utilization by end users serving specific applications, while the latter provides a formal representation of concepts acting as a common reference point for aggregating data about the entire healthcare enterprise. An example application facilitated by an interface terminology is to support clinicians' entry of patient information into an EHR.

To create an initial CIT, we integrated concepts from all existing COVID ontologies in BioPortal [4]. We used the operations of concatenation and anchoring to mine candidate chunks from the clinical text of COVID-19 patients. These chunks were reviewed by a domain expert and accepted chunks were included in the CIT to further enrich its conceptual content. Besides being useful in their own right, these chunks mined from EHR notes of COVID-19 patients can serve as training data for machine learning techniques, enabling further large-scale concept mining. Having a system in place that can quickly capture clinically relevant snippets from clinical notes, recognize those that can be annotated with existing concepts from a dedicated Interface Terminology, and incorporating new concepts rapidly into the Interface Terminology will be immensely useful. This would support discovery and research on emerging diseases such as COVID-19.

II. BACKGROUND

A. Interface terminologies

An Interface Terminology [3, 5] is designed to maximize the utilization by end users of a specific application, e.g., facilitating the direct entry of clinical data into EHR systems. Since interface terminologies are designed with user-specific applications in mind, they usually contain colloquial usages and common clinical phrases constituting a richer synonym content compared to reference terminologies. One of the recommendations for developing an interface terminology is to construct it from an existing ontology [6]. We follow this approach by constructing the initial version of our COVID interface terminology by integrating existing COVID ontologies/terminologies.

B. Existing COVID-19 Ontologies/Terminologies

COVID-19 research is a fast-moving area. Several public ontologies have recently been built that incorporate COVID concepts or are completely dedicated to it and are accessible through the NCBO BioPortal [4]. Foremost on BioPortal is the Coronavirus Infectious Disease Ontology (CIDO) [7] which was created with the aim to provide a standardized representation of various coronavirus infectious diseases. CIDO ensures interoperability by following the OBO Foundry [8] principles and integrating concepts from about 20 other ontologies like ChEBI [9], HPO [10], etc. More than 60 new concepts related to COVID-19 have also been codified in CIDO. CIDO covers multiple areas in the domain of coronavirus diseases (etiology, diagnosis, treatment, etc.), and these various components are linked by relations, e.g., *caused by* and *treatment for*. Drug concepts are mapped from ChEBI, NDF-RT [11], and DrON [12], and different forms of the coronavirus, e.g., *Coronavirus Neoromicia* are mapped from NCBI-Taxonomy [13]. As of September 30, CIDO contains 6009 classes (concepts) and is rapidly growing.

The COVID-19 ontology [14] (2268 concepts) predominantly covers concepts related to cell types, genes, and proteins involved in virus-host-interactions, as well as medical and epidemiological concepts relevant to COVID-19. This ontology is similar to CIDO, but includes more concepts related to diseases affecting various systems of the human body. The COVID-19 Infectious Disease Ontology (IDO-COVID-19) [15] (486 concepts) extends the Infectious Disease Ontology (IDO) [16] and the Virus Infectious Disease Ontology (VIDO) [17] to solely represent concepts related to the virus and diseases associated with COVID-19. The World Health Organization's (WHO) COVID-19 Rapid Version CRF semantic data model (COVIDCRFRAPID) [18] (398 concepts) aims at capturing the semantic references to the questions and answers in the case report form. Apart from this, there are two small ontologies in BioPortal - the COviD-19 Ontology for Cases and Patient information (CODO) [19], and the COVID-19 Surveillance Ontology (COVID19) [20], both with 52 concepts and mainly dealing with concepts related to the surveillance, geography, treatment facilities and tracking of patients.

The ACT COVID Ontology v3.0 is available on GitHub [21] as SQL files that can be loaded into a database and was created to support cohort identification and related research by incorporating terms related to diagnosis, procedure, and medication codes from ICD [22], LOINC [23], CPT [24] and NDC [25]. We extracted 2,446 concepts from the available files. In addition to these, UMLS, SNOMED CT [26], and LOINC have published lists of concepts related to COVID-19 on their respective websites.

C. Dataset – Radiology Case Studies

EHRs contain sensitive health information. While dissemination of EHR notes could benefit clinical trials and controlled studies during this pandemic it must be done in a way that ensures patient privacy. The privacy rule of HIPAA [27] outlines policies for sharing personal data and the mandatory de-identification processes are rigorous and time-consuming. Due to these reasons, large-scale de-identified EHRs are not yet publicly available. Since we were not able to find unstructured clinical text related to COVID patients, we used an alternative, public data source discussed below.

Radiology imaging reports, including Computed Tomography (CT) imaging and X-rays play a major role in the diagnosis and management of COVID-19 patients. Radiology case studies are chronicles of patient progress describing classic and unusual presentations of diseases with a focus on findings in CTs and X-rays. For this study, we used 115 COVID-19 radiology case studies from the Italian Society of Radiology [28]. These studies describe patient demographic information such as age and sex, prior medical history, symptoms, detailed CT and chest X-ray findings, and medications.

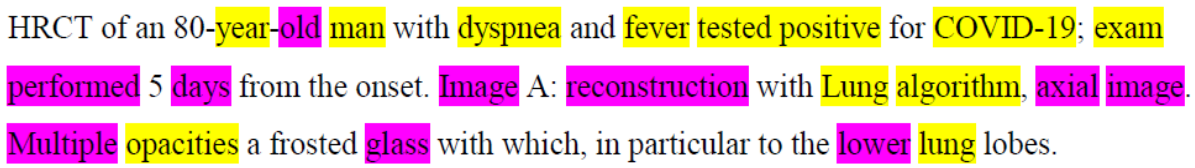

Fig. 1. A snippet from a radiology case study of a COVID-19 patient annotated with ICIT (in yellow) and DIFF (in pink).

D. Annotation Tools

A suite of biomedical annotation tools is available for use. Some of these are general purpose tools in that they detect a wide variety of entity types and in most cases links terms to concepts in the UMLS. The cTAKES [29], MetaMap [30], QuickUMLS [31], and NCBO Annotator [32, 33] are examples of general purpose annotation tools. Another category of annotation tools is trained to identify specific entity types like genes, disease, chemicals, etc. PubTator [34] and BERN [35] are examples in this category. Yet another category of annotation tools is trained on manually annotated datasets using machine learning techniques and can work on a wide variety of entity types based on the training data available. CLAMP [36] is an example of such an annotation tool.

III. METHODS

A. Creating an Initial COVID Interface Terminology (ICIT)

We created an Initial COVID Interface Terminology (ICIT) by including into CIDO the concepts from the five COVID-related ontologies in BioPortal mentioned in section II.B (COVID-19, CODO, COVIDCRFRAPID, COVID19, IDO-COVID-19), as well as the ACT COVID ontology and COVID-related concepts from SNOMED CT, LOINC, and the UMLS. Due to its size, continuous growth, and abundance of available relationships, CIDO was selected as the backbone for ICIT, into which concepts from the other ontologies were integrated. During the integration process, we removed duplicates, which occurred, because several of the included ontologies reused identical concepts from other terminologies, e.g., from ChEBI and NDF-RT. For example, there were 481 concepts common in CIDO and COVID-19. Apart from this, we also found concepts in different ontologies with different concept names that are synonyms of each other. We collected such concepts under a single concept ID with a unique concept name and assigned other concept names as synonyms of the concept.

For the experiments reported in this paper we created a separate hierarchy for the concepts from each of the different ontologies integrated into CIDO, since for testing the annotation process, only the list of concepts is needed and not their arrangement in a hierarchy. However, our terminal goal is to integrate all these concepts according to their relationships into a coherent COVID Interface Terminology hierarchy (concept network).

B. NCBO Annotator as the Annotation Tool

We used the NCBO annotator (NCBOA) [32] for annotating the radiology case studies in our dataset. This decision was based on the following requirements of this study. The first requirement is that we needed a general purpose annotation tool. Since the purpose is to annotate clinical text and capture all relevant medical information, we do not want to restrict ourselves to annotators that are trained on specific entity types. For example, PubTator cannot recognize procedures (e.g. laparoscopy). Secondly, we do not have manually annotated training data supporting concept mining for COVID-19. Hence state-of-the-art tools like CLAMP cannot be used. Finally, we needed the ability to annotate text with concepts from a sequence of versions of the interface terminology that we created.

Even though annotator tools such as cTAKES, MetaMap, and QuickUMLS were found to perform better than NCBOA, these tools use the UMLS or source vocabularies from the UMLS for the purpose of annotation. NCBOA provides the unique advantage that the annotation can be performed with any custom terminology uploaded by a user into BioPortal. Furthermore, the existing COVID terminologies that we enumerated in II.B are all present in BioPortal. Thus, we uploaded our ICIT terminology into BioPortal and used it with NCBOA to annotate the radiology text samples.

C. Extracting Auxiliary Concepts

Apart from specific disease-related information and medications, EHRs also contain the anamnesis of a patient, which has a huge role in deciding the course of the treatment. For example, the sentence "*74 year old male with Phmx of nephrolithiasis, prostate ca s/p XRT presented to ED…*" extracted from a synthetic COVID clinical note provided by (AE), describes the prior history of kidney stones and prostate cancer. Such concepts are not in ICIT. However, they are essential for the annotation of EHRs of COVID-19 patients. Thus, we want to add such concepts to CIT under the appropriate auxiliary hierarchies. SNOMED CT is a good source for extracting such concepts.

We created a program (DIFF) to identify these auxiliary concepts. The program outputs a list of annotated concepts by identifying the difference in the annotations between the text annotated with SNOMED CT and with ICIT. In other words, DIFF identifies all the concepts annotated with SNOMED CT that are not present in ICIT. Fig. 1 shows an excerpt from a radiology case study of a COVID-19 patient annotated with ICIT (in yellow) and the DIFF (in pink). The concepts such as *old*, *reconstruction*, and *lower* are examples of auxiliary concepts. We applied DIFF to the collection of 115 radiology case studies in our dataset to collect all auxiliary concepts and integrated them into ICIT to form a new interface terminology, the CIT Version 0 (CIT_v0).

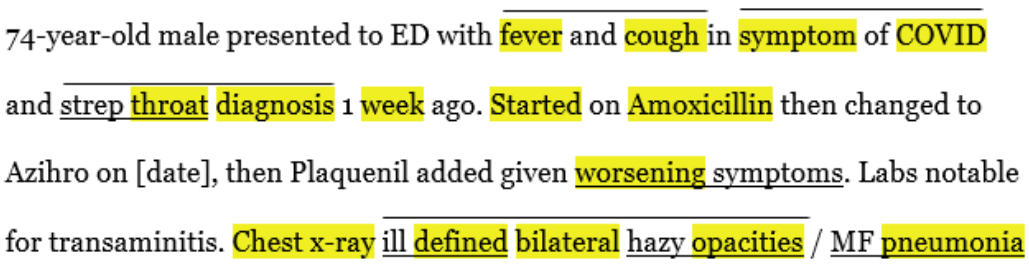

Fig. 2. An excerpt from a synthetic EHR illustrating some example phrases obtained by concatenation and anchoring procedures. Overbars represent concatenation and underlines represent anchoring.

D. *Mining Concepts for building the COVID Interface Terminology*

As discussed before, reference terminologies often do not contain many of the high granularity phrases that appear in EHR notes. Because of this, some critical information is lost during the annotation process. This issue can be addressed by mining concepts from the EHR itself. Thus, extracting high granularity concepts from EHR notes is one of the challenges to overcome for enriching CIT with such essential concepts.

For addressing this challenge, we used concatenation and anchoring of existing concepts as follows. **Concatenation** involves combining two or more adjacent existing concepts into a high granularity phrase. We allow stop words in between. **Anchoring** extracts phrases by adding one or two words to the left, right or both sides of an existing concept, and we allow stop words to intervene. For example, consider w1, w2 as two words, sw as a stop word, and define * to mean 0, 1, or more occurrences [Kleene Star in Algorithms], then the candidate anchoring phrases can be represented using the following three rules. The "+" stands for string concatenation.

1. w1 + sw* + [existing concept]
2. [existing concept] + sw* + w1
3. w1 + sw* + [existing concept] + sw* + w2

We will illustrate these two techniques using Fig. 2, which is an excerpt from a synthetic note of a COVID-19 patient provided by (AE) and annotated with CIT_v0. Concatenation is marked by overbars and anchoring is marked by underlines. For example, the existing concepts *symptom* and *COVID* can be concatenated to form the chunk *symptom of COVID*. The concept *strep throat* is obtained by anchoring "strep" to the existing concept *throat*. In the next step, this new concept will then be concatenated with *diagnosis* providing the chunk *strep throat diagnosis*. Similarly, we obtain the chunk *ill defined bilateral hazy opacities*, by first applying anchoring to get *ill defined* and *hazy opacities* and then by concatenating these phrases with *bilateral*.

To mine chunks, we applied the concatenation and anchoring procedures alternatingly on our dataset, annotating it with CIT_v0. To be explicit, we first annotated text with CIT_v0 concepts. Then we applied concatenation. Those phrases that were accepted by a human expert were then added to CIT_v0 to obtain CIT_v1.1. Next, we annotated the dataset with CIT_v1.1 and applied anchoring to obtain more candidate phrases. The phrases accepted by the expert were added to CIT_v1.1 to obtain CIT_v1.2. This process continued, alternating between concatenation and anchoring. The advantage of alternating the concatenation and anchoring steps is that the phrases obtained by concatenation can participate in anchoring in the next step and vice versa. For example, *strep throat* mentioned above is obtained as a concept in CIT_v1.2 by anchoring and is used in the subsequent concatenation phase with the concept *diagnosis* to obtain the phrase *strep throat diagnosis* as a concept in CIT_v2.1.

Since concatenation and anchoring are brute-force techniques, human review is necessary. Thus, after each application of concatenation and anchoring the extracted phrases were reviewed in a two step process. Concepts were first prescreened by the core team and then the accepted candidates were reviewed by a medical expert. Prescreening was possible, because the majority of automatically generated phrases were parts of larger chunks or spanned two partial chunks. For example, “thickening of pulmonary” is a part of *thickening of pulmonary interstitium,* and “MRSA and port-a-cath” spans two chunks, *sepsis from MRSA* and *port-a-cath infection*. All the phrases that passed both review steps were integrated into the version of the CIT that was current at that time.

We note that all the phrases that were rejected at any review step were automatically excluded from the candidate phrases list and never appeared again in the subsequent processing steps. Hence, each rejected phrase is reviewed only once, saving review time. Similarly, the accepted phrases were integrated into the CIT as concepts and used for annotation in the next iteration. Thus they cannot appear again as candidate phrases. Therefore, the number of extracted phrases decreases significantly after each application of concatenation and anchoring. After a few iterations, when the number of new phrases falls below a threshold, the processing is terminated.

After the domain expert review, we performed a synonym check on the accepted phrases. Phrases that are synonyms are combined under a single concept ID. One phrase is chosen as the concept name and the other phrases are labeled as synonyms. This is exemplified by the two phrases *history positive for contact with COVID-19 patient* and *positive history of contact with COVID-19 patient*.

As a safeguard against false negatives at the first review step (by the core team), we created a sample of 200 phrases, selected randomly from the rejected phrases in all the

TABLE I. STATISTICS OF EXTRACTED PHRASES FOR ALL VERSIONS

Version	Procedure	Total # phrases	# phrases after 1st review	% phrases after 1st review	# phrases after 2nd review	% phrases after 2nd review	% retained w.r.t 1st review
CIT_v1.1	Concatenation	1893	873	46.12%	781	41.25%	89.5%
CIT_v1.2	Anchoring	3923	1590	40.53%	1351	34.44%	84.97%
CIT_v2.1	Concatenation	1002	439	43.81%	389	38.82%	88.6%
CIT_v2.2	Anchoring	969	295	30.44%	268	27.66%	90.86%
CIT_v3.1	Concatenation	314	92	29.30%	83	26.43%	90.20%
CIT_v3.2	Anchoring	185	34	18.37%	30	16.21%	88.24%
CIT_v4.1	Concatenation	66	6	9.09%	6	9.09%	100%
CIT_v4.2	Anchoring	69	6	8.69%	6	8.69%	100%

iterations. This sample was then reviewed by the domain expert to check for cases of false negatives, i.e., acceptable phrases.

E. Evaluation metrics

We evaluated the performance of our techniques using two metrics – Coverage and Breadth. **Coverage** is the percentage of words being annotated. **Breadth** is the average number of words per annotated concept. Using chunks rather than concepts in reference terminologies increases breadth.

$$Coverage = (Number\ of\ words\ in\ all\ annotated\ concepts \div Total\ number\ of\ words) * 100 \quad (1)$$

$$Breadth = Number\ of\ words\ in\ all\ annotated\ concepts \div Number\ of\ annotated\ concepts \quad (2)$$

For example, annotating our dataset with ICIT, we obtained 2330 annotated concepts with a total of 2845 words out of the 20,994 words in the dataset. Thus the coverage is 13.55% and the breadth is 1.22.

IV. RESULTS

A. ICIT and CIT_v0

To create the ICIT, we integrated concepts from other COVID terminologies into CIDO. After removing duplicates, 1780 concepts from COVID-19 were added into CIDO, as well as 352 concepts from IDO-COVID-19, 272 concepts from COVIDCRFRAPID, 46 concepts from CODO, and 50 concepts from the COVID-19 Surveillance Ontology. From the ACT COVID ontology a total of 2445 concepts were included. In addition to this, we incorporated 113, 74, and 2 concepts from SNOMED, LOINC, and UMLS, respectively. After identifying and accounting for synonyms, the total number of concepts in ICIT at this stage was 10,024.

For creating the CIT_v0, we integrated auxiliary concepts from SNOMED CT into ICIT, as discussed in III.C. We identified 904 auxiliary concepts. Thus the total number of concepts in CIT_v0 increased to 10,928.

B. Different versions of CIT

To mine new chunks for the CIT, we annotated our dataset with NCBOA & CIT by alternatingly applying concatenation and anchoring. The accepted phrases after review by the expert were added to the previous version of CIT to obtain the subsequent version. This was expressed in III.D by incrementing version numbers. The numbers of extracted phrases, the numbers of phrases retained after the core team (1st) reviews and after the expert (2nd) reviews, and the corresponding percentages are in Table I for all the versions of CIT created thus far. The last column of the table shows the percentages of phrases that were retained by the expert with respect to the percentages from the core team review.

In Table II, we show examples of phrases obtained as a result of concatenation during the creation of CIT_v1.1. The existing concepts that were combined to form the chunks are shown between two '|' symbols. The phrase *tested positive for COVID-19* was obtained by combining two existing concepts *tested positive* and *COVID-19,* allowing for the stop word "for." Another example, *history of contact with Covid-19 patient* is a combination of four existing concepts, as shown in the third row in Table II.

In Table II, we also show examples obtained by applying anchoring that were accepted for inclusion in CIT_v1.2. The existing concept that was used as an anchor is marked in bold. The phrase *subpleural distribution* is an example of the first rule of anchoring, where a left word was added to the existing concept *distribution*. The phrase, *mediastinal lympha-denomegalies* is the result of applying the second rule of anchoring, which adds a right word; *ground glass areas* demonstrates the third rule of combining both left and right words with an existing concept.

TABLE II. EXAMPLES OF ACCEPTED PHRASES FROM CIT_V1

Version	Accepted Phrases
CIT_v1.1 (concatenation)	|tested positive| for |COVID-19|
	|history of| |contact with| |Covid-19| |patient|
	|peri|-|bronchial| |thickening|
	|limited| |lymphadenopathy|
	|spider| |web| |sign|
CIT_v1.2 (anchoring)	subpleural **distribution**
	mediastinal lymphadenomegalies
	parenchymal **thickening**
	interstitial-alveolar **pneumonia**
	ground **glass** areas

Examples of rejected phrases are shown in Table III for both concatenation and anchoring. As in Table II, for concatenation the existing concepts are between two '|' symbols and for anchoring they are marked in bold. As discussed in Methods (III.D), there are phrases that are part of longer chunks (e.g., partial pleurogenic) or spanning two chunks (e.g., axis with Fogarty catheter).

TABLE III. EXAMPLES OF REJECTED PHRASES

Procedure	Rejected Phrases
Concatenation	|hyperpyrexia| |refractory|
	|axis| with |Fogarty catheter|
	|thrombocytopenia| and |need for|
	|chest x-ray| with |multiple|
	|pneumonia| with |radiographic|
Anchoring	**Multiple opacities** a frosted
	increased density with a ground
	segment of the **upper**
	Partial pleurogenic
	reticular and interstitial

Examples of longer phrases that were added to CIT_v2 and CIT_v3 are provided in Table IV. For example, the phrase *extensive areas with crazy-paving patterns* was obtained for addition to CIT_v2.1 as a result of concatenating two phrases *extensive areas* and *crazy-paving patterns* that were already in CIT_v1.1. Similarly, the phrase *subpleural distribution* present in CIT_v1.2 was used as an anchor to extract *predominantly subpleural distribution* for inclusion in CIT_v2.2.

TABLE IV. EXAMPLES OF ACCEPTED LONGER PHRASES

Accepted Longer Phrases
inter- and intra-lobular septal thickening
extensive areas with crazy-paving patterns
parenchymal consolidation area with subpleural distribution
bilateral subpleural ground glass opacities
widespread fibrotic-like reticular bands
parenchymal consolidations in both upper lobes
predominantly subpleural distribution
centrolobular and subpleural paraseptal emphysema

To check for false negatives among the phrases that were rejected by the core team, we created a random sample of 200 phrases from all the iterations. This sample was reviewed by the domain expert. The domain expert found only eight out of the 200 phrases to be false negatives.

C. Evaluation metrics

We annotated our dataset with CIDO, ICIT, and CIT_v0, obtaining coverages of 6%, 13.55%, and 40.84%, respectively. The breadths were 1.18, 1.22, and 1.21, respectively. The coverage and breadth for different versions of CIT are shown in Table V. The number of concepts in each version is also shown.

TABLE V. COVERAGE AND BREADTH

Version	# concepts	Coverage	Breadth
CIT_v1.1	11,644	41.30%	1.55
CIT_v1.2	12,984	53.66%	2.16
CIT_v2.1	13,364	53.97%	2.47
CIT_v2.2	13,628	58.09%	2.65
CIT_v3.1	13,711	58.19%	2.73
CIT_v3.2	13,741	58.41%	2.74
CIT_v4.1	13,747	58.42%	2.74
CIT_v4.2	13,753	58.46%	2.74

V. DISCUSSION

The approach presented in this paper is based on several assumptions. The first assumption is that by mining concepts from clinical notes of COVID-19 patients and including them in the interface terminology, CIT will contain many concepts that correspond to cognitive chunks used by MDs. Numerous such concepts do not appear in reference terminologies. However, we observe that the multi-word names of such concepts tend to contain shorter concepts that do appear in the reference terminologies. Thus, we can use the operations of anchoring and concatenation of existing concepts in the CIT_v0 to obtain higher granularity concepts that correspond to chunks used by MDs.

The higher granularity concept candidates are obtained algorithmically after annotating the dataset with CIT_v0. The only manual step that requires a domain expert is reviewing which of the generated phrases obtained by concatenation or anchoring are valid for inclusion in the CIT. As was discussed in the Methods section, each phrase is reviewed only once during the life cycle of creating the CIT. Hence, despite the manual review required, the creation of the interface terminology is efficient. Once the interface terminology has been created, it can be used for annotation of an unlimited number of clinical notes of COVID-19 patients. The process of generating chunks comprises the alternating application of concatenation and anchoring, such that one operation provides the extended input for the other operation.

The second assumption is that the described process will converge. That means that on average later steps will find fewer and fewer new phrases than those preceding them. We assume that although individual MDs write free text in individualized ways, the number of possible chunks used in a specific discipline, in this case, COVID-19 is fundamentally limited and is growing slower as the field matures. First of all, different MDs might use similar, but not identical, phrases to express the same concept. One of those will be designated the name of the concept and all the others will be synonyms of the concept in CIT. When annotating text with CIT, the synonyms are also identified and annotated. A second reason for assuming a limited number of chunks in each discipline is that the length of the chunks is limited to a few words. Thus, we will not encounter an exponential growth in the number of new terms, but more likely a polynomial increase, which is considered manageable in the theory of computing.

Our current dataset is too small to check the above assumption. We plan a larger experiment where we expect to see the convergence of the numbers of new chunks obtained. The current study introduces the operations of concatenation and anchoring for this domain, and the successful iterative, alternating use of these operations to create a more comprehensive interface terminology for COVID-19. Our results show convergence, with the number of extracted and accepted phrases decreasing with each iteration. In the last version, CIT_v4 we accepted only 12 phrases for both concatenation and anchoring combined.

As mentioned before, the results of concatenation and anchoring need to be reviewed by a human expert. The time of domain experts is a limited and expensive resource. To minimize the use of this resource, our team members, who are not MDs, performed a preliminary review, after having gone through training based on samples that were previously reviewed by domain experts. The main purpose of this preliminary review was to exclude phrases that obviously should not be part of the CIT. Only the phrases accepted in the preliminary review were passed on to the domain expert for validation. According to Table I, for the first six versions of CIT, on average about 88% of the phrases accepted by the core team were also validated by the domain expert. Hence it is safe to say that an initial review by the core team helped to eliminate a large number of phrases that were not corresponding to concepts, thereby minimizing the time and effort expended by the domain expert.

The domain expert reviewed a sample of 200 phrases that had been rejected by the core team and found only 4% of them to be viable for inclusion in the CIT. We also observed a consistent decrease in the percentages of phrases accepted by the reviews of the results of concatenation and anchoring (Table I). For concatenation, the acceptance percentage for the core team review decreased from 46.12% to 43.81% to 29.30% to 9.09%, and for expert review from 41.25% to 38.82% to 26.43% to 9.09%. Acceptance rates for anchoring follow the same trend with (40.53%→30.44%→18.37%→8.69%) for core team review and (34.44%→27.66%→16.21%→8.69%) for expert review. This can be attributed to the many illegal combinations arising when the phrases became longer with each iteration and that were pruned in the process.

The concatenation and anchoring operations have a different impact on the evaluation metrics. Concatenation provides a minimum contribution to the coverage of the dataset. This is because concatenation combines already existing annotated concepts and hence does not add new words to the annotated word list, except for the stop words that bridge the gaps between existing concepts. There is only a 0.31% change in coverage obtained by concatenation in CIT_v2.1 compared to the coverage of the previous version CIT_v1.2. Similarly, the change in coverage is only 0.10% when moving from version CIT_v2.2 to CIT_v3.1. Concatenation favors breadth since breadth increases with the length of the phrases, and combining existing phrases increases the length.

Anchoring tends to increase both coverage and breadth. During anchoring "unannotated" words are added to the left, right, or on both sides of a concept, and hence accepted phrases resulting from anchoring capture words that were previously not annotated and contribute to increasing the number of annotated words, thereby increasing coverage. Anchoring also increases breadth, as the newly added words increase the length of the phrases annotated by specific concepts. The increases in coverage obtained by the first three anchoring iterations are 12.36%, 4.12%, and 0.22%. The corresponding increases in breadth are 0.61, 0.18, and 0.01.

We have selected the COVID-19 dataset to support ongoing research for medications and vaccinations for COVID-19. However, the methodology described for designing an interface terminology to support annotation of EHR is applicable for other medical specialties assuming the design of a dedicated interface terminology for individual specialty.

Future work: As mentioned, this work describes a feasibility study with a small dataset. In the future, we will conduct an extensive study that will test the assumption that after a while the chunks appearing in new clinical notes start repeating and already exist in the interface terminology, and only a very few new concepts are added in later iterations (convergence). For this study, we will randomly divide the dataset into two parts, the "training" set and the test set. The interface terminology will be created based on the training set and then it will be used to annotate the test set.

Our hypothesis is that the coverage and breadth values for the test dataset will be marginally smaller than that for the sample set, but almost on par, which would provide one data point to demonstrate the generalizability of our approach.

For the extensive study with a larger dataset, we plan to only review phrases that appear more than once in the dataset. These phrases are more likely to appear in the test dataset than those which appear only once. For the extensive study, the number of phrases that appear only once is likely to be too overwhelming to afford reviews. Moreover, many of them are not likely to appear in the test dataset. Furthermore, we hypothesize that the phrases with higher frequency in the dataset have higher acceptance rates. To test this hypothesis, we will study the correlation between frequency and acceptance. The dataset used in the current study is not large enough to ignore the phrases which appear only once.

VI. CONCLUSIONS

Unstructured clinical text from Electronic Health Records (EHRs) contains valuable information about patient progress, and when annotated properly could help advance research on emerging infectious diseases such as COVID-19. In this paper, we exhibited the design of an interface terminology (CIT) for the annotation of clinical notes of COVID-19 patients.

The interface terminology was initialized with concepts from several COVID ontologies and with existing general purpose concepts (non-COVID concepts) from SNOMED CT encountered in the dataset. Its content was significantly extended by mining high granularity concepts from these

clinical notes. We introduced the operations of concatenation and anchoring for this research and applied them alternatingly and iteratively. Version 4.2 of the CIT (CIT_v4.2) achieved a 43% increase in coverage compared to CIT_v0 that the iteration process started with.

ACKNOWLEDGMENT

Research reported in this publication was supported by the National Center for Advancing Translational Sciences (NCATS), a component of the National Institute of Health (NIH) under award number UL1TR003017. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

[1] Miller GA. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol. Rev. 1956;63:81.
[2] Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Res. 2004;32:D267-D70.
[3] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J. Am. Med. Inform. Assoc. 2006;13:277-88.
[4] Whetzel P, Shah N, Noy N, Dai B, Dorf M, Griffith N, et al. BioPortal: Ontologies and integrated data resources at the click of a mouse. Nature Precedings. 2009.
[5] Kanter AS, Wang AY, Masarie FE, Naeymi-Rad F, Safran C. Interface terminologies: bridging the gap between theory and reality for Africa. Stud. Health Technol. Inform. 2008:27-32.
[6] Rosenbloom ST, Brown SH, Froehling D, Bauer BA, Wahner-Roedler DL, Gregg WM, et al. Using SNOMED CT to Represent Two Interface Terminologies. J. Am. Med. Inform. Assoc. 2009;16:81-8.
[7] He Y, Yu H, Ong E, Wang Y, Liu Y, Huffman A, et al. CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Scientific Data. 2020;7:1-5.
[8] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 2007;25:1251-5.
[9] Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2007;36:D344-D50.
[10] Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics. 2008;83:610-5.
[11] National Drug File - Reference Terminology Source Information. https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/index.html, (accessed Sept 30 2020).
[12] Hanna J, Joseph E, Brochhausen M, Hogan WR. Building a drug ontology based on RxNorm and other sources. Journal of biomedical semantics. 2013;4:44.
[13] Federhen S. The NCBI taxonomy database. Nucleic Acids Res. 2012;40:D136-D43.
[14] COVID-19 Ontology. http://bioportal.bioontology.org/ontologies/COVID-19, (accessed Sept 30 2020).
[15] Babcock S, Cowell LG, Beverley J, Smith B. The Infectious Disease Ontology in the Age of COVID-19. 2020.
[16] Infectious Disease Ontology. https://bioportal.bioontology.org/ontologies/IDO, (accessed Sept 30 2020).
[17] Virus Infectious Disease Ontology. https://bioportal.bioontology.org/ontologies/VIDO, (accessed Sept 30 2020).
[18] WHO COVID-19 Rapid Version CRF semantic data model. https://bioportal.bioontology.org/ontologies/COVIDCRFRAPID, (accessed Sept 30 2020).
[19] Dutta B, DeBellis M. CODO: An Ontology for Collection and Analysis of Covid-19 Data. arXiv preprint arXiv:2009.01210. 2020.
[20] de Lusignan S, Bernal JL, Zambon M, Akinyemi O, Amirthalingam G, Andrews N, et al. Emergence of a novel coronavirus (COVID-19): protocol for extending surveillance used by the Royal College of general practitioners research and surveillance centre and public health England. JMIR public health and surveillance. 2020;6:e18606.
[21] ACT COVID Ontology v3.0. https://github.com/shyamvis/ACT-COVID-Ontology/tree/master/ontology, (accessed Sept 30 2020).
[22] WHO. International Classification of Diseases. http://www.who.int/classifications/icd/en/, (accessed Sept 30 2020).
[23] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin. Chem. 2003;49:624-33.
[24] Hirsch JA, Leslie-Mazwi TM, Nicola GN, Barr RM, Bello JA, Donovan WD, et al. Current procedural terminology; a primer. J. Neurointerv. Surg. 2015;7:309-12.
[25] National Drug Code Database Background Information. https://www.fda.gov/drugs/development-approval-process-drugs/national-drug-code-database-background-information, (accessed Sept 30 2020).
[26] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud. Health Technol. Inform. 2006;121:279-90.
[27] Annas GJ. HIPAA regulations—a new era of medical-record privacy? : Mass Medical Soc; 2003.
[28] SIRM. COVID-19 Database. https://www.sirm.org/category/senza-categoria/covid-19/, (accessed Jun 5 2020).
[29] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc. 2010;17:507-13.
[30] Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J. Am. Med. Inform. Assoc. 2010;17:229-36.
[31] Soldaini L, Goharian N. Quickumls: a fast, unsupervised approach for medical concept extraction. MedIR workshop, sigir2016. p. 1-4.
[32] Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit on Translat Bioinforma.;2009:56-60.
[33] Tchechmedjiev A, Abdaoui A, Emonet V, Melzi S, Jonnagaddala J, Jonquet C. Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator+. Bioinformatics. 2018;34:1962-5.
[34] Wei C-H, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019;47:W587-W93.
[35] Kim D, Lee J, So CH, Jeon H, Jeong M, Choi Y, et al. A neural named entity recognition and multi-type normalization tool for biomedical text mining. IEEE Access. 2019;7:73729-40.
[36] Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J. Am. Med. Inform. Assoc. 2018;25:331-6.