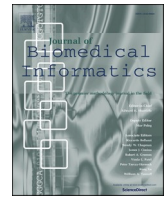




Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework

Kevin Lybarger^{a,*}, Mari Ostendorf^b, Matthew Thompson^c, Meliha Yetisgen^a

^a Biomedical & Health Informatics, University of Washington, Box 358047, Seattle, WA 98109, USA

^b Department of Electrical & Computer Engineering, University of Washington, Campus Box 352500 185, Seattle, WA 98195-2500, USA

^c Department of Family Medicine, University of Washington, Box 354696, Seattle, WA 98195-2500, USA

ARTICLE INFO

Keywords:

COVID-19

Coronavirus

Machine learning

Natural language processing

Information extraction

ABSTRACT

Coronavirus disease 2019 (COVID-19) is a global pandemic. Although much has been learned about the novel coronavirus since its emergence, there are many open questions related to tracking its spread, describing symptomatology, predicting the severity of infection, and forecasting healthcare utilization. Free-text clinical notes contain critical information for resolving these questions. Data-driven, automatic information extraction models are needed to use this text-encoded information in large-scale studies. This work presents a new clinical corpus, referred to as the COVID-19 Annotated Clinical Text (CACT) Corpus, which comprises 1,472 notes with detailed annotations characterizing COVID-19 diagnoses, testing, and clinical presentation. We introduce a span-based event extraction model that jointly extracts all annotated phenomena, achieving high performance in identifying COVID-19 and symptom events with associated assertion values (0.83–0.97 F1 for events and 0.73–0.79 F1 for assertions). Our span-based event extraction model outperforms an extractor built on MetaMapLite for the identification of symptoms with assertion values. In a secondary use application, we predicted COVID-19 test results using structured patient data (e.g. vital signs and laboratory results) and automatically extracted symptom information, to explore the clinical presentation of COVID-19. Automatically extracted symptoms improve COVID-19 prediction performance, beyond structured data alone.

1. Introduction

As of December 20, 2020, there were over 75 million confirmed COVID-19 cases globally, resulting in 1.6 million related deaths [1]. Surveillance efforts to track the spread of COVID-19 and estimate the true number of infections remains a challenge for policy makers, healthcare workers, and researchers, even as testing availability increases. Symptom information provides useful indicators for tracking potential COVID-19 infections and disease clusters [2]. Certain symptoms and underlying comorbidities have directed COVID-19 testing. However, the clinical presentation of COVID-19 varies significantly in severity and symptom profiles [3].

The most prevalent COVID-19 symptoms reported to date are fever, cough, fatigue, and dyspnea [4], but emerging reports identify additional symptoms, including diarrhea and neurological symptoms, such as changes in taste or smell [5–7]. Certain initial symptoms may be associated with higher risk of complications; in one study, dyspnea was associated with a two-fold increased risk of acute respiratory distress

syndrome [8]. However, correlations between symptoms, positive tests, and rapid clinical deterioration are not well understood in ambulatory care and emergency department settings.

Routinely collected information in the Electronic Health Record (EHR) can provide crucial COVID-19 testing, diagnosis, and symptom data needed to address these knowledge gaps. Laboratory results, vital signs, and other structured data results can easily be queried and analyzed at scale; however, more detailed and nuanced descriptions of COVID-19 diagnoses, exposure history, symptoms, and clinical decision-making are typically only documented in the clinical narrative. To leverage this textual information in large-scale studies, the salient COVID-19 and symptom information must be automatically extracted.

This work presents a new corpus of clinical text annotated for COVID-19, referred to as the COVID-19 Annotated Clinical Text (CACT) Corpus. CACT consists of 1,472 notes from the University of Washington (UW) clinical repository with detailed event-based annotations for COVID-19 diagnosis, testing, and symptoms. The event-based annotations characterize these phenomena across multiple dimensions,

* Corresponding author.

E-mail address: lybarger@uw.edu (K. Lybarger).

including assertion, severity, change, and other attributes needed to comprehensively represent these clinical phenomena in secondary use applications. This work is part of a larger effort to use routinely collected data describing the clinical presentation of acute and chronic diseases, with two major aims; (1) to describe the presence, character, and changes in symptoms associated with clinical conditions, where delays or misdiagnoses occur in clinical practice and impact patient outcomes (e.g. infectious diseases, cancer) [9], and (2) to provide a more efficient and cost-effective mechanism to validate clinical prediction rules previously derived from large prospective cohort studies [10]. To the best of our knowledge, CACT is the first clinical data set with COVID-19 annotations, and it includes 29.9K distinct events. We present the first information extraction results on CACT using an end-to-end neural event extraction model, establishing a strong baseline for identifying COVID-19 and symptom events. We explore the prediction of COVID-19 test results (positive or negative) using structured EHR data and automatically extracted symptoms and find that the automatically extracted symptoms improve prediction performance.

2. Related work

2.1. Annotated corpora

Given the recent onset of COVID-19, there are limited COVID-19 corpora for natural language processing (NLP) experimentation. Corpora of scientific papers related to COVID-19 are available [11,12], and automatic labels for biomedical entity types are available for some of these research papers [13]. However, we are unaware of corpora of clinical text with supervised COVID-19 annotations.

Multiple clinical corpora are annotated for symptoms. As examples, South et al. [14] annotated symptoms and other medical concepts with negation (present/not present), temporality, and other attributes. Koeling et al. [15] annotated a pre-defined set of symptoms related to ovarian cancer. For the i2b2/VA challenge, Uzun et al. [16] annotated medical concepts, including symptoms, with assertion values and relations. While some of these corpora may include symptom annotations relevant to COVID-19 (e.g. “cough” or “fever”), the distribution and characterization of symptoms in these corpora may not be consistent with COVID-19 presentation. To fill the gap in clinical COVID-19 annotations and detailed symptom annotation, we introduce CACT to provide a relatively large corpus with COVID-19 diagnosis, testing, and symptom annotations.

2.2. Medical concept and symptom extraction

The most commonly used United Medical Language System (UMLS) concept extraction systems are the clinical Text Analysis and Knowledge Extraction System (cTAKES) [17] and MetaMap [18]. The National Library of Medicine (NLM) created a lightweight Java implementation of MetaMap, *MetaMapLite*, which demonstrated real-time speed and extraction performance comparable to or exceeding the performance of MetaMap, cTAKES, and DNorm [19]. In previous work, we built on *MetaMapLite*, incorporating assertion value predictions (e.g. present versus absent) using classifiers trained on the 2010 i2b2 challenge dataset to create the extraction pipeline referred to here as *MetaMapLite++* [20]. *MetaMapLite++* assigns each extracted UMLS Metathesaurus concept an assertion value with an Support Vector Machine (SVM)-based assertion classifier that utilizes syntactic and semantic knowledge. The SVM assertion classifier achieved state-of-the-art assertion performance (Micro-F1 94.23) on the i2b2 2010 assertion dataset [21]. Here, we use *MetaMapLite++* as a baseline for evaluating extraction performance for a subset of our annotated phenomena,

specifically symptoms with assertion values, using the UMLS “Sign or Symptom” semantic type. The Mayo Clinic updated its rule-based medical tagging system, MedTagger [22], to include a COVID-19 specific module that extracts 18 phenomena related to COVID-19, including 11 common COVID-19 symptoms with assertion values [23]. We do not use the COVID-19 MedTagger variant as a baseline, because our symptom annotation and extraction is not limited to known COVID-19 symptoms.

2.3. Relation and event extraction

There is a significant body of information extraction (IE) work related to coreference resolution, relation extraction, and event extraction tasks. In these tasks, spans of interest are identified, and linkages between spans are predicted. Many contemporary IE systems use end-to-end multi-layer neural models that encode an input word sequence using recurrent or transformer layers, classify spans (entities, arguments, etc.), and predict the relationship between spans (coreference, relation, role, etc.) [24–29]. Of most relevance to our work is a series of developments starting with Lee et al. [30], which introduces a span-based coreference resolution model that enumerates all spans in a word sequence, predicts entities using a feed-forward neural network (FFNN) operating on span representations, and resolves coreferences using a FFNN operating on entity span-pairs. Luan et al. [31] adapts this framework to entity and relation extraction, with a specific focus on scientific literature. Luan et al. [32] extends the method to take advantage both of co-reference and relation links in a graph-based approach to jointly predict entity spans, co-reference, and relations. By updating span representations in multi-sentence co-reference chains, the graph-based approach achieved state-of-the-art on several IE tasks representing a range of different genres. Wadden et al. [33] expands on Luan et al. [32]’s approach, adapting it to event extraction tasks. We build on Luan et al. [31] and Wadden et al. [33]’s work, augmenting the modeling framework to fit the CACT annotation scheme. In CACT, event arguments are generally close to the associated trigger, and inter-sentence events linked by co-reference are infrequent, so the graph-based extension, which adds complexity, is unlikely to benefit our extraction task.

Many recent NLP systems use pre-trained language models (LMs), such as ELMo, BERT, and XLNet, that leverage unannotated text [34–36]. A variety of strategies for incorporating the LM output are used in IE systems, including using the contextualized word embedding sequence: as the input to a Conditional Random Field entity extraction layer [37], as the basis for building span representations [32,33], or by adding an entity-aware attention mechanism and pooled output states to a fully transformer-based model [38]. There are many domain-specific LM variants. Here, we use Alsentzer et al. [39]’s *Bio+Clinical BERT*, which is trained on PubMed papers and MIMIC-III [40] clinical notes.

2.4. COVID-19 outcome prediction

There are many pre-print and published works exploring the prediction of COVID-19 outcomes, including COVID-19 infection, hospitalization, acute respiratory distress syndrome, need for intensive care unit (ICU), need for a ventilator, and mortality [41–53]. These COVID-19 outcomes are typically predicted using existing structured data within the EHR, including demographics, diagnosis codes, vitals, and lab results, although Izquierdo et al. [46] incorporates automatically extracted information from the existing EHRead tool. Our literature review identified 24 laboratory, vital sign, and demographic fields that are predictive of COVID-19 (see Table 7 in the Appendix details). While there are some frequently cited fields, there does not appear to be a consensus across the literature regarding the most prominent predictors

Table 1

Annotation guideline summary. * indicates the argument is required. † indicates at least one of the arguments, *Test Status* or *Assertion*, is required.

Event type, <i>e</i>	Argument type, <i>a</i>	Argument subtypes, L_i	Span examples
COVID	Trigger*	–	“COVID,” “COVID-19”
	Test Status†	{positive, negative, pending, conditional, not ordered, not patient, indeterminate}	“tested positive”
	Assertion†	{present, absent, possible, hypothetical, not patient}	“positive,” “low suspicion”
Symptom	Trigger*	–	“cough,” “shortness of breath”
	Assertion*	{present, absent, possible, conditional, hypothetical, not patient}	“admits,” “denies”
	Change	{no change, worsened, improved, resolved}	“improved,” “continues”
	Severity	{mild, moderate, severe}	“mild,” “required ventilation”
	Anatomy	–	“chest wall,” “lower back”
	Characteristics	–	“wet productive,” “diffuse”
	Duration	–	“for two days,” “1 week”
	Frequency	–	“occasional,” “chronic”

of COVID-19 infection. These 24 predictive fields informed the development of our COVID-19 prediction work in Section 5. Prediction architectures include logistic regression, SVM, decision trees, random forest, K-nearest neighbors, Naïve Bayes, and multilayer perceptron [46,47,51–53].

3. Materials

3.1. Data

This work used inpatient and outpatient clinical notes from the UW clinical repository. COVID-19-related notes were identified by searching for variations of the terms *coronavirus*, *covid*, *sars-cov*, and *sars-2* in notes authored between February 20–March 31, 2020, resulting in a pool of 92K notes. Samples were randomly selected for annotation from a subset of 53K notes that include at least five sentences and correspond to the note types: telephone encounters, outpatient progress, emergency department, inpatient nursing, intensive care unit, and general inpatient medicine. Multiple note types were used to improve extraction model generalizability.

Early in the outbreak, the UW EHR did not include COVID-19 specific structured data; however, structured fields indicating COVID-19 test types and results were added as testing expanded. We used these structured fields to assign a *COVID-19 Test* label describing COVID-19 polymerase chain reaction (PCR) testing to each note based on patient test status within the UW system (no data external to UW was used):

- *none*: patient testing information is not available
- *positive*: patient will have at least one future positive test
- *negative*: patient will only have future negative tests

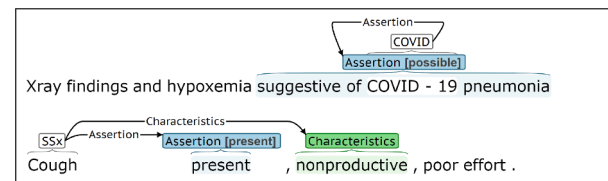


Fig. 1. BRAT annotation examples for *COVID* and *Symptom* (SSx) event types.

More nuanced descriptions of COVID-19 testing (e.g. conditional or unordered tests) or diagnoses (e.g. possible infection or exposure) are not available as structured data. For the 53K note subset, the *COVID-19 Test* label distribution is 90.8% *none*, 7.9% *negative*, and 1.3% *positive*.¹

Given the sparsity of *positive* and *negative* notes, CACT is intentionally biased to increase the prevalence of these labels. To ensure adequate *positive* training samples, the CACT training partition includes 46% *none*, 5% *negative*, and 49% *positive* notes. Ideally, the test set would be representative of the true distribution; however, the expected number of *positive* labels with random selection is insufficient to evaluate extraction performance. Consequently, the CACT test partition was biased to include 50% *none*, 46% *negative*, and 4% *positive* notes. Notes were randomly selected in equal proportions from the six note types. CACT includes 1,472 annotated notes, including 1,028 train and 444 test notes.

3.2. Annotation scheme

We created detailed annotation guidelines for COVID-19 and symptoms, using the event-based annotation scheme in Table 1. Each event includes a trigger that identifies and anchors the event and arguments that characterize the event. The annotation scheme includes two types of arguments: *labeled arguments* and *span-only arguments*. *Labeled arguments* (e.g. *Assertion*) include an argument span, type, and subtype (e.g. *present*). The subtype label normalizes the span information to a fixed set of classes and allows the extracted information to be directly used in secondary use applications. *Span-only arguments* (e.g. *Characteristics*) include an argument span and type but do not include a subtype label, because the argument information is not easily mapped to a fixed set of classes.

For *COVID* events, the trigger is generally an explicit COVID-19 reference, like “COVID-19” or “coronavirus.” *Test Status* characterizes implicit and explicit references to COVID-19 testing, and *Assertion* captures diagnoses and hypothetical references to COVID-19. *Symptom* events capture subjective, often patient reported, indications of disorders and diseases (e.g. “cough”). For *Symptom* events, the trigger identifies the specific symptom, for example “wheezing” or “fever,” which is characterized through *Assertion*, *Change*, *Severity*, *Anatomy*, *Characteristics*, *Duration*, and *Frequency* arguments. Symptoms were annotated for all conditions/diseases, not just COVID-19. Notes were annotated using the BRAT annotation tool [54]. Fig. 1 presents BRAT annotation examples.

Most prior medical problem extraction work, including symptom extraction, focuses on identifying the specific problem, normalizing the extracted phenomenon, and predicting an assertion value (e.g. present versus absent). This approach omits many of the symptom details that clinicians are taught to document and that form the core of many clinical notes. This symptom detail describes change (e.g. improvement, worsening, lack of change), severity (e.g. intensity and impact on daily activities), particular characteristics (e.g. productive, dry, or barking for cough), and location. We hypothesize that this symptom granularity is

¹ The COVID-19 test positivity rate cannot be inferred from these label distributions, as there can be multiple test results associated with each note-level label.

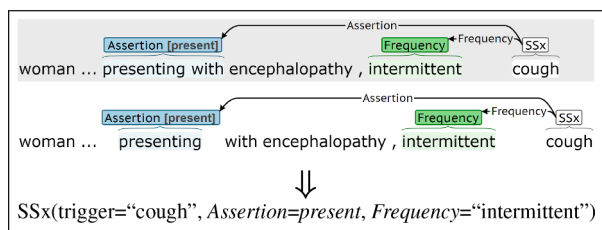


Fig. 2. Annotation examples describing event extraction as a slot filling task.

needed for many clinical conditions to improve timely diagnosis and validate diagnosis prediction rules.

3.3. Annotation scoring and evaluation

Annotation and extraction is scored as a slot filling task, focusing on information most relevant to secondary use applications. Fig. 2 presents the same sentence annotated by two annotators, along with the populated slots for the *Symptom* event. Both annotations include the same trigger and *Frequency* spans (“cough” and “intermittent”, respectively). The *Assertion* spans differ (“presenting with” vs. “presenting”), but the assigned subtypes (*present*) are the same, so the annotations are equivalent for purposes of populating a database. Annotator agreement and extraction performance are assessed using scoring criteria that reflects this slot filling interpretation of the labeling task.

The *Symptom* trigger span identifies the specific symptom. For *COVID*, the trigger anchors the event, although the span text is not salient to downstream applications. For labeled arguments, the subtype label captures the most salient argument information, and the identified span is less informative. For span-only arguments, the spans are not easily mapped to a fixed label set, so the selected span contains the salient information. Performance is evaluated using precision (P), recall (R), and F1.

Trigger: Triggers, T_i , are represented by a pair (event type, e_i ; token indices, x_i). Trigger equivalence is defined as

$$T_i \equiv T_j \text{ if } (e_i \equiv e_j) \wedge (x_i \equiv x_j).$$

Arguments: Events are aligned based on trigger equivalence. The arguments of events with equivalent triggers are compared using different criteria for *labeled arguments* and *span-only arguments*. Labeled arguments, L_i , are represented as a triple (argument type, a_i ; token indices, x_i ; subtype, l_i). For labeled arguments, the argument type, a , and subtype, l , capture the salient information and equivalence is defined as

$$L_i \equiv L_j \text{ if } (T_i \equiv T_j) \wedge (a_i \equiv a_j) \wedge (l_i \equiv l_j).$$

Span-only arguments, S_i , are represented as a pair (argument type, a_i ; token indices, x_i). Span-only arguments with equivalent triggers and argument types, $(T_i \equiv T_j) \wedge (a_i \equiv a_j)$, are compared at the token-level (rather than the span-level) to allow partial matches. Partial match scoring is used as partial matches can still contain useful information.

3.4. Annotation statistics

CACT includes 1,472 notes with a 70%/30% train/test split and 29.9K annotated events (5.4K *COVID* and 24.4K *Symptom*). Fig. 3 contains a summary of the *COVID* annotation statistics for the train/test subsets. By design, the training and test sets include high rates of COVID-19 infection (*present* subtype for *Assertion* and *positive* subtype for *Test Status*), with higher rates in the training set. CACT includes high rates of *Assertion hypothetical* and *possible* subtypes. The *hypothetical* subtype applies to sentences like, “She is mildly concerned about the coronavi-

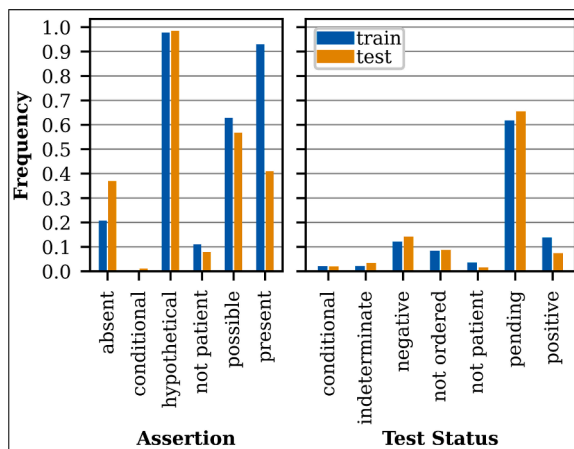


Fig. 3. COVID annotation summary.

rus” and “She cancelled nexplanon replacement due to COVID-19.” The *possible* subtype applies to sentences like, “risk of Covid exposure” and “Concern for respiratory illness (including COVID-19 and influenza).” *Test Status pending* is also frequent.

There is some variability in the endpoints of the annotated *COVID* trigger spans (e.g. “COVID” vs. “COVID test”); however 98% of the *COVID* trigger spans in the training set start with the tokens “COVID,” “COVID19,” or “coronavirus.” Since the *COVID* trigger span is only used to anchor and disambiguate events, the *COVID* trigger spans were truncated to the first token of the annotated span in all experimentation and results.

The training set includes 1,756 distinct uncased *Symptom* trigger spans, 1,425 of which occur fewer than five times. Fig. 4 presents the frequency of the 20 most common *Symptom* trigger spans in the training set by *Assertion* subtypes *present*, *absent*, and other (*possible*, *conditional*, *hypothetical*, or *not patient*). The extracted symptoms in Fig. 4 were manually normalized to aggregate different extracted spans with similar meanings (e.g. “sob” and “short of breath” → “shortness of breath”; “febrile” and “fevers” → “fever”). Table 6 in the Appendix presents the

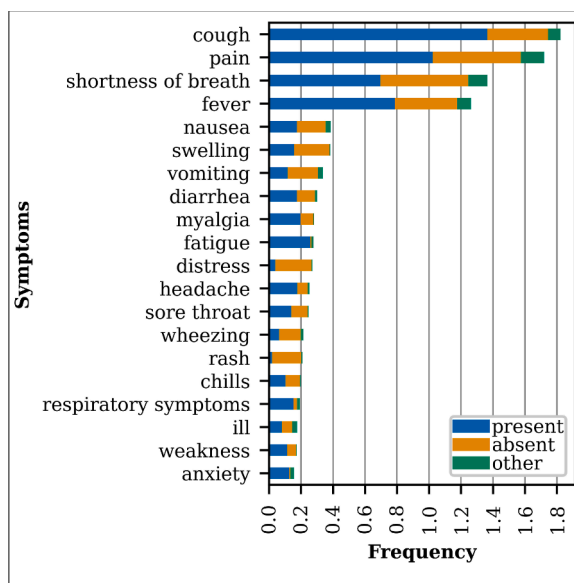


Fig. 4. Most frequent symptoms in the training set broken down by *Assertion* subtype.

symptom normalization mapping, provided by a medical doctor. These 20 symptoms account for 62% of the training set *Symptom* events. There is ambiguity in delineating between some symptoms and other clinical phenomena (e.g. exam findings and medical problems), which introduces some annotation noise.

Given the long tail of the symptom distribution and our desire to understand the more prominent COVID-19 symptoms, we focused annotator agreement assessment and extraction model training/evaluation on the symptoms that occurred at least 10 times in the training set, resulting in 185 distinct, unnormalized symptoms that cover 82% of the training set *Symptom* events. The set of 185 symptoms was determined only using the training set, to allow unbiased experimentation on the test set. The subsequent annotator agreement and information extraction experimentation only incorporate these 185 most frequent symptoms.

3.5. Annotator agreement

All annotation was performed by four UW medical students in their fourth year. After the first round of annotation, annotator disagreements were carefully reviewed, the annotation guidelines were updated, and annotators received additional training. Additionally, potential COVID triggers were pre-annotated using pattern matching (“COVID,” “COVID-19,” “coronavirus,” etc.), to improve the recall of COVID annotations. Pre-annotated COVID triggers were modified as needed by the annotators, including removing, shifting, and adding trigger spans. Fig. 5 presents the annotator agreement for the second round of annotation, which included 96 doubly annotated notes. For labeled arguments, F1 scores are micro-average across subtypes.

4. Event extraction

4.1. Methods

Event extraction tasks, like ACE05 [55], typically require prediction of the following event phenomena:

- trigger span identification
- trigger type (event type) classification
- argument span identification
- argument type/role classification

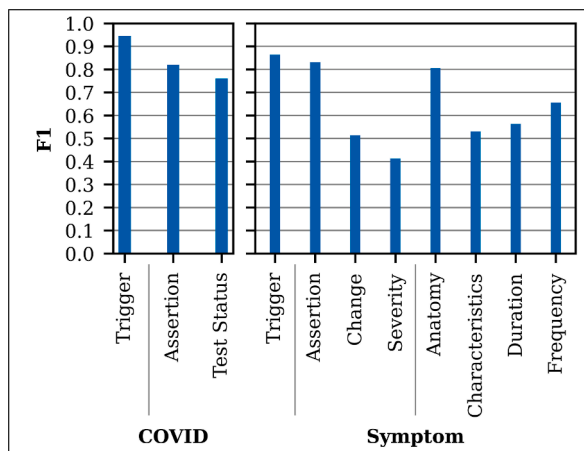


Fig. 5. Annotator agreement.

The CACT annotation scheme differs from this configuration in that labeled arguments require the argument type (e.g. *Assertion*) and the subtype (e.g. *present*, *absent*, etc.) to be predicted. Resolving the argument subtypes require a classifier with additional predictive capacity.

We implement a span-based, end-to-end, multi-layer event extraction model that jointly predicts all event phenomena, including the trigger span, event type, and argument spans, types, and subtypes. Fig. 6 presents our Span-based Event Extractor framework, which differs from prior related work in that multiple span classifiers are used to accommodate the argument subtypes.

Each input sentence consists of tokens, $X = \{x_1, x_2, \dots, x_n\}$, where n is the number of tokens. For each sentence, the set of all possible spans, $S = \{s_1, s_2, \dots, s_m\}$, is enumerated, where m is the number of spans with token length less than or equal to M tokens. The model generates trigger and argument predictions for each span in S and predicts the pairing between arguments and triggers to create events from individual span predictions.

Input encoding: Input sentences are mapped to contextualized word embeddings using *Bio+Clinical BERT* [39]. To limit computational cost, the contextualized word embeddings feed into a bi-LSTM without fine tuning BERT (no backpropagation to BERT). The bi-LSTM has hidden size v_h . The forward and backward states, $h_{t,f}$ and $h_{t,b}$, are concatenated to form the $1 \times 2v_h$ dimensional vector $h_t = [h_{t,f}, h_{t,b}]$, where t is the token position.

Span representation: Each span is represented as the attention weighted sum of the bi-LSTM hidden states. Separate attention mechanisms, c , are implemented for trigger and each labeled argument, and a single attention mechanism is implemented for all span-only arguments, $c \in \{1, 2, \dots, 6\}$ (1 for trigger, 4 for labeled arguments, and 1 for span-only arguments). The attention score for span representation c at token position t is calculated as

$$\alpha_{c,t} = w_{a,c} h_t^T \tag{1}$$

where $w_{a,c}$ is a learned $1 \times 2v_h$ vector. For span representation c , span i , and token position t , the attention weights are calculated by normalizing the attention scores as

$$a_{c,i,t} = \frac{\exp(\alpha_{c,t})}{\sum_{k=start(s_i)}^{end(s_i)} \exp(\alpha_{c,k})} \tag{2}$$

where $start(s_i)$ and $end(s_i)$ denote the start and end token indices of span s_i . Span representation c for span i is calculated as the attention-

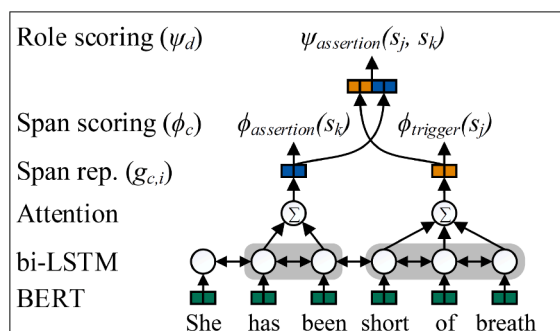


Fig. 6. Span-based Event Extractor.

Table 2
Extraction performance.

Event type	Argument	Train-CV				Test			
		# Gold	P	R	F1	# Gold	P	R	F1
COVID	Trigger	3,931	0.95	0.97	0.96	1,497	0.96	0.97	0.97
	Assertion	2,936	0.70	0.74	0.72	1,075	0.72	0.74	0.73
	Test Status	1,068	0.60	0.62	0.61	457	0.63	0.60	0.62
Symptom	Trigger	13,823	0.82	0.85	0.83	5,789	0.81	0.85	0.83
	Assertion	13,833	0.77	0.79	0.78	5,791	0.77	0.80	0.79
	Change	739	0.45	0.03	0.06	341	0.45	0.05	0.09
	Severity	743	0.47	0.30	0.37	327	0.45	0.31	0.37
	Anatomy	3,839	0.76	0.59	0.66	1,959	0.78	0.50	0.61
	Characteristics	3,145	0.59	0.26	0.36	1,441	0.66	0.25	0.36
	Duration	3,744	0.62	0.44	0.51	1,344	0.54	0.56	0.55
	Frequency	801	0.64	0.39	0.48	250	0.60	0.51	0.55

weighted sum of the bi-LSTM hidden state as

$$\mathbf{g}_{c,i} = \sum_{t=\text{start}(s_i)}^{\text{end}(s_i)} a_{c,i,t} \mathbf{h}_t. \quad (3)$$

Span prediction: Similar to the span representations, separate span classifiers, c , are implemented for trigger and each labeled argument, and a single classifier predicts all span-only arguments, $c \in \{1, 2, \dots, 6\}$ (1 for trigger, 4 for labeled arguments, and 1 for span-only arguments). Label scores for classifier c and span i are calculated as

$$\phi_c(s_i) = \mathbf{w}_{s,c} \text{FFNN}_{s,c}(\mathbf{g}_{c,i}), \quad (4)$$

where $\phi_c(s_i)$ yields a vector of label scores of size $|L_c|$, $\text{FFNN}_{s,c}$ is a non-linear projection from size $2v_h$ to v_s , and $\mathbf{w}_{s,c}$ has size $|L_c| \times v_s$. The trigger prediction label set is $L_{\text{trigger}} = \{\text{null}, \text{COVID}, \text{Symptom}\}$. Separate classifiers are used for each labeled argument (*Assertion*, *Change*, *Severity*, and *Test Status*) with label set, $L_c = \{\text{null} \cup L_l\}$, where L_l is defined in Table 1.² For example, $L_{\text{Severity}} = \{\text{null}, \text{mild}, \text{moderate}, \text{severe}\}$. A single classifier predicts all span-only arguments with label set, $L_{\text{span-only}} = \{\text{null}, \text{Anatomy}, \text{Characteristics}, \text{Duration}, \text{Frequency}\}$.

Argument role prediction: The argument role layer predicts the assignment of arguments to triggers using separate binary classifiers, d , for each labeled argument and one classifier for all span-only arguments, $d \in \{1, 2, \dots, 5\}$ (4 for labeled arguments and 1 for span-only arguments). Argument role scores for trigger j and argument k using argument role classifier d are calculated as

$$\psi_d(s_j, s_k) = \mathbf{w}_{r,d} \text{FFNN}_{r,d}([g_j, g_k]) \quad (5)$$

where $\psi_d(s_j, s_k)$ is a vector of size 2, $\text{FFNN}_{r,d}$ is a non-linear projection from size $4v_h$ to v_r , and $\mathbf{w}_{r,d}$ has size $2 \times v_r$.

Span pruning: To limit the time and space complexity of the pairwise argument role predictions, only the top- K spans for each span classifier, c , are considered during argument role prediction. The span score is calculated as the maximum label score in ϕ_c , excluding the *null* label score.

4.2. Model Configuration

The model configuration was selected using 3-fold cross validation (CV) on the training set. Table 8 in the Appendix summarizes the selected configuration. Training loss was calculated by summing the cross entropy across all span and argument role classifiers. Models were implemented using the Python PyTorch module [56].

4.3. Data representation

During initial experimentation, *Symptom Assertion* extraction performance was high for the *absent* subtype and lower for *present*. The higher *absent* performance is primarily associated with the consistent presence of negation cues, like “denies” or “no.” While there are affirming cues, like “reports” or “has,” the *present* subtype is often implied by a lack of negation cues. For example, an entire sentence could be “Short of breath.” To provide the *Symptom Assertion* span classifier with a more consistent span representation, we replaced each *Symptom Assertion* span (token indices) with the *Symptom* trigger span in each event and found that performance improved. We extended this trigger span substitution approach to all labeled arguments (*Assertion*, *Change*, *Severity*, and *Test Status*) and found performance improved. By substituting the trigger spans for the labeled argument spans, trigger and labeled argument prediction is roughly treated as a multi-label classification problem, although the model does not constrain trigger and labeled argument predictions to be associated with the same spans. As previously discussed, the scoring routine does not consider the span indices of labeled arguments.

4.4. Results

Table 2 presents the event extraction performance on the training set using CV and the withheld test set. Extraction performance is similar on the train and test sets, even though the training set has higher rates of COVID-19 positive notes. *COVID* trigger extraction performance is very high (0.97 F1) and comparable to the annotator agreement (0.95 F1). The *COVID Assertion* performance (0.73 F1) is higher than *Test Status* performance (0.62 F1), which is likely due to the more consistent *Assertion* annotation. *Symptom* trigger and *Assertion* extraction performance is high (0.83 F1 and 0.79 F1, respectively), approaching the annotator agreement (0.86 F1 and 0.83 F1, respectively). *Anatomy* extraction performance (0.61 F1) is lower than expected, given the high annotator agreement (0.81 F1). *Duration* extraction performance is comparable to annotator agreement, and *Frequency* extraction performance is lower than annotation agreement. *Change*, *Severity*, and *Characteristics* extraction performance is low, again likely related to low annotator agreement for these cases.

Existing symptom extraction systems do not extract all of the phenomena in the CACT annotation scheme; however, we compared the performance of our Span-based Event Extractor to MetaMapLite++ for symptom identification and assertion prediction. MetaMapLite++ is an analysis pipeline that includes a UMLS concept extractor with assertion prediction [20]. Table 3 presents the performance of MetaMapLite++ on the CACT test set. The spans associated with medical concepts in MetaMapLite++ differ slightly from our annotation scheme. For example, “dry cough” was extracted by MetaMapLite++ as a symptom,

² The assertion classifier uses the larger label set associated with *Symptom*.

Table 3
MetaMapLite++ extraction performance for *Symptom trigger* and *Assertion*.

Case	Argument	P	R	F1
Exact trigger match	Trigger	0.53	0.54	0.54
	Assertion	0.43	0.44	0.44
Any triggers overlap	Trigger	0.66	0.67	0.66
	Assertion	0.54	0.55	0.54

whereas our annotation scheme labels “cough” as the symptom and “dry” as a characteristic. To account for this difference, Table 3 presents the performance of MetaMapLite++ for two trigger equivalence criteria: (1) *exact match* for triggers is required, as defined in Section 3.3 and 2) *any overlap* for triggers is considered equivalent. The Span-based Event Extractor outperforms MetaMapLite++ for symptom identification precision (0.81 vs. 0.66), recall (0.85 vs. 0.67), and F1 (0.83 vs. 0.66), even when MetaMapLite++ is evaluated with the more relaxed *any overlap* trigger scoring. The lower recall of MetaMapLite++ is partially the result of the UMLS not including symptom acronyms and abbreviations that frequently occur in our data, for example “N/V/D” for “nausea, vomiting, and diarrhea.” Table 3 only reports the performance for the UMLS “Sign or Symptom” semantic type. When all UMLS semantic types are used, the recall improves; however, the precision is extremely low ($P = 0.02 - 0.03$)³.

Table 4 presents the assertion prediction performance for both systems, only considering the subset of predictions with exact trigger matches (i.e. assertion prediction performance is assessed without incurring penalty for trigger identification errors). The number of gold assertion labels (“# Gold”) is greater for the Span-based Event Extractor, because more of the symptom triggers predictions are correct. The Span-based Event Extractor outperforms MetaMapLite++ in assertion prediction precision (0.95 vs. 0.81), recall (0.94 vs. 0.81), and F1 (0.94 vs. 0.81). MetaMapLite++’s lower performance is partly the result of differences between the distribution of assertion labels in CACT and the dataset used to train MetaMapLite++’s assertion classifier (2010 i2b2).

5. COVID-19 prediction application

The creation of the CACT Corpus and the Span-based Event Extractor is motivated by our larger effort to explore the clinical presentation of diseases through the comprehensive representation of symptoms across multiple dimensions. This section utilizes a subset of the extracted information to predict positive COVID-19 infection status among individuals presenting to clinical settings for COVID-19 testing. This experimentation uses COVID-19 test results to distinguish between COVID-19 negative and COVID-19 positive patients with the goals of identifying the clinical presentation of COVID-19 and investigating the predictive power of symptoms. Improved understanding of the clinical presentation of COVID-19 has the potential to improve risk stratification of patients presenting for COVID-19 testing (by increasing or decreasing their pre-test probability), and thus guide diagnostic testing and clinical decision making.

5.1. Data

An existing clinical data set from the UW from January 2020 through May 2020 was used to explore the prediction of COVID-19 test results and identify the most prominent predictors of COVID-19. The data set represents 230K patients, including 28K patients with at least one COVID-19 PCR test result. The data set includes telephone encounters,

³ In the UMLS, 15% of the unique gold symptoms in the CACT training set are covered when only the “Sign or Symptom” semantic type is used. The UMLS coverage increases to 48% when all semantic types are used, and the unique gold symptoms occur in 76 different UMLS semantic types.

Table 4
Symptom Assertion comparison for events with equivalent triggers (exact span match).

Model	# Gold	P	R	F1
MetaMap++	3,152	0.81	0.81	0.81
Span-based Event Extractor	4,952	0.95	0.94	0.94

outpatient progress notes, and emergency department (ED) notes, as well as structured data (demographics, vitals, laboratory results, etc.).

For each patient in this data set, all of the COVID-19 tests with either a *positive* or *negative* result and at least one note within the seven days preceding the test result were identified. Only COVID-19 tests with a note within the previous seven days are included in experimentation, to improve the robustness of the COVID-19 symptomology exploration. Each of these test results was treated as a sample in this binary classification task (positive or negative). The likelihood of COVID-19 positivity was predicted using structured EHR data and notes within a 7-day window preceding the test result. The pairing of notes and COVID-19 test results was independently performed for each of the note types (ED, outpatient progress, and telephone encounter notes). From this pool of data, we identified the following test counts by note type: 2,226 negative and 148 positive for ED; 7,599 negative and 381 positive for progress; and 7,374 negative and 448 positive for telephone. Within the 7-day window of this subset of COVID-19 test results, there are 5.3K ED, 14.5K progress, and 27.5K telephone notes. This data set has some overlap with the data set used in Section 4.1 but is treated as a separate data set in this COVID-19 prediction task. The notes in the CACT training set are less than 1% of the notes used in this secondary use application.

5.2. Methods

Features: Symptom information was automatically extracted from the notes using the Span-based Event Extractor trained on CACT.⁴ The extracted symptoms were normalized using the mapping in Table 6 in the Appendix. Each extracted symptom with an *Assertion* value of “present” was assigned a feature value of 1. The 24 identified predictors of COVID-19 from existing literature (see Section 2) were mapped to 32 distinct fields within the UW EHR and used in experimentation. Identified fields are listed in Table 9 of the Appendix. For the coded data (e.g. structured fields like “basophils”), experimentation was limited to this subset of literature-supported COVID-19 predictors, given the limited number of positive COVID-19 tests in this data set.

Within the 7-day history, features may occur multiple times (e.g. multiple temperature measurements). For each feature, the series of values was represented as the minimum or maximum of the values depending on the specific feature. For example, temperature was represented as the maximum of the measurements to detect any fever, and oxygen saturation was represented as the minimum of the values to capture any low oxygenation events. Table 9 in the Appendix includes the aggregating function, f , used for each field.

Where symptom features were missing, the feature value was set to 0. For features from the structured EHR data, which are predominantly numerical, missing features were assigned the mean feature value in the set used to train the COVID-19 prediction model.

Model: COVID-19 was predicted using the Random Forest framework, because it facilitates nonlinear modeling with interdependent features and interpretability analyses (Scikit-learn Python implementation used [57]). Alternative prediction algorithms include Logistic Regression, SVM, and FFNN. Logistic Regression assumes feature independence and linearity, which is not valid for this task. For example, the feature set includes both the symptom “fever” and temperature

⁴ Only automatically extracted symptom data were used. No supervised (hand annotated) labels were used.

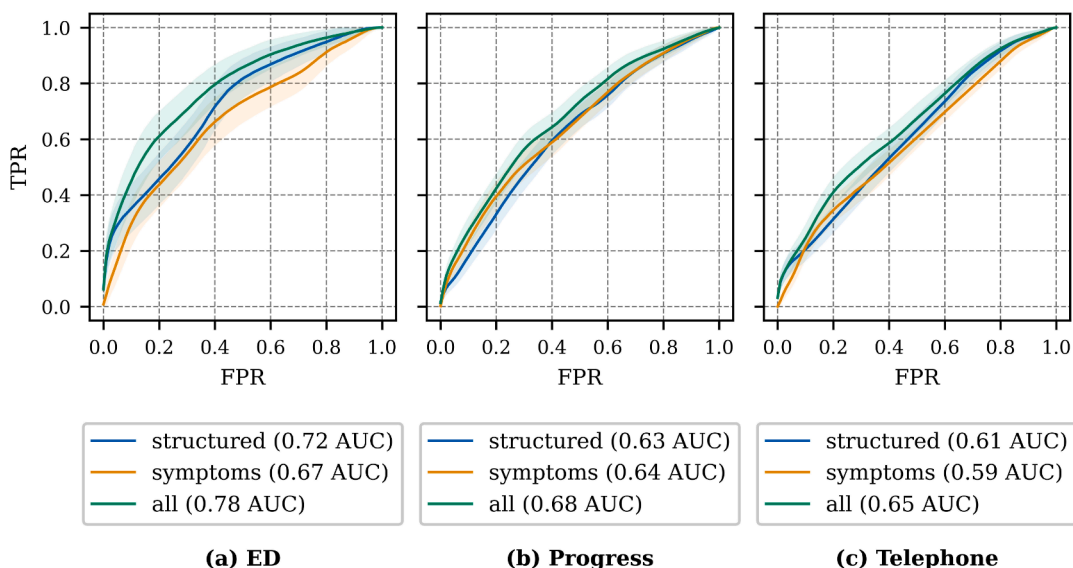


Fig. 7. Receiver operating characteristic by note type and feature set combination for repeated hold-out iterations. The solid line indicates the average ROC, and the shaded region around the solid line indicates one standard deviation.

measurements (e.g. “38.6°C”). Model interpretability is less clear with SVM, and the number of positive test samples is relatively small for a FFNN.

The relative importance of features in predicting COVID-19 was explored using Lundberg et al. [58]’s SHAP (SHapley Additive exPlanations) approach, which is implemented in the SHAP Python module.⁵ SHAP generates interpretable, feature-level explanations for nonlinear model predictions. For each prediction, SHAP feature scores are estimated, where larger absolute scores indicate higher importance, and the absolute values of the scores sum to 1.0 for each prediction.

Experimental paradigm: The available data was split into train/test sets using an 80%/20% split by patient, although training and evaluation was performed at the test-level (i.e. each COVID-19 test result is a sample). Performance was evaluated using the receiver operating characteristic (ROC) and the associated area under the curve (AUC). Given the relatively small number of positive samples, the train/test splits were randomly created 1,000 times through repeated hold-out testing [59]. Kim [59] demonstrated that repeated hold-out testing can improve the robustness of the results in low resource settings. For each train/test split, the AUC was calculated, and an average AUC was calculated across all hold-out iterations. The random holdout iterations yield a distribution of AUC values, which facilitate significance testing. The significance of the AUC performance was assessed using a two-sided T-test. The Random Forest models were tuned using 3-fold cross validation on the training set and evaluated on the withheld test set. COVID-19 prediction experimentation included three feature sets: *structured* (32 structured EHR fields), *notes* (automatically extracted symptoms), and *all* (combination of structured fields and automatically extracted symptoms). Separate models were trained and evaluated for each note type (ED, progress, and telephone) and feature set (*structured*, *notes*, and *all*). The selected Random Forest hyperparameters are summarized in Table 10 in the Appendix.

5.3. Results

Fig. 7 presents the ROC for the COVID-19 predictors with the average AUC across repeated hold-out partitions. The AUC evaluates model performance across all operating points, including operating points that

are not clinically significant, for example extremely low true positive rate (TPR). To address this AUC limitation and provide an alternative method for comparing feature sets, we selected a fixed operating point on the ROC, comparing the false positive rate (FPR) at a specific TPR. We selected a TPR (sensitivity) of 80%, as a value that has clinical value for identifying individuals with COVID-19, and we examined the FPR (specificity) at this fixed TPR. In this use case, we are attempting to see how well structured EHR fields and symptoms perform compared to the reference standard of a laboratory PCR test.

Table 5 presents the FPR at TPR=0.80, including the FPR mean and standard deviation across the repeated holdout iterations. Lower FPRs (better performance) are achieved for all three note types, when automatically extracted symptoms are added to the structured data. We would not expect a combination of clinical features to have particularly high sensitivity. Smith et al. [60] achieved similar performance in predicting COVID-19 using clinical prediction rules. While detecting COVID-19 in 80% of patients with the disease, the inclusion of automatically extracted symptoms decreases the FPR (the “cost”) by 2-7 percentage points. For all note types, the inclusion of the automatically extracted symptom information (*all* feature set) improves performance over structured data only (*structured-only* feature set) for both AUC and FPR@TPR=0.80 with significance ($p < 0.001$ per two-sided T-test). The *structured* features achieve higher performance in the ED note experimentation, than experimentation with progress and telephone notes, due to the higher prevalence of vital sign measurements and

Table 5
COVID-19 prediction false positive rate at a true positive rate of 80%.

Note type	Feature type	FPR @ TPR = 0.80
ED	structured	0.48 ± 0.09
	symptoms	0.63 ± 0.12
	all	0.41 ± 0.10
Progress	structured	0.64 ± 0.05
	symptoms	0.64 ± 0.05
	all	0.58 ± 0.06
Telephone	structured	0.66 ± 0.04
	symptoms	0.71 ± 0.03
	all	0.64 ± 0.05

⁵ <https://pypi.org/project/shap/>

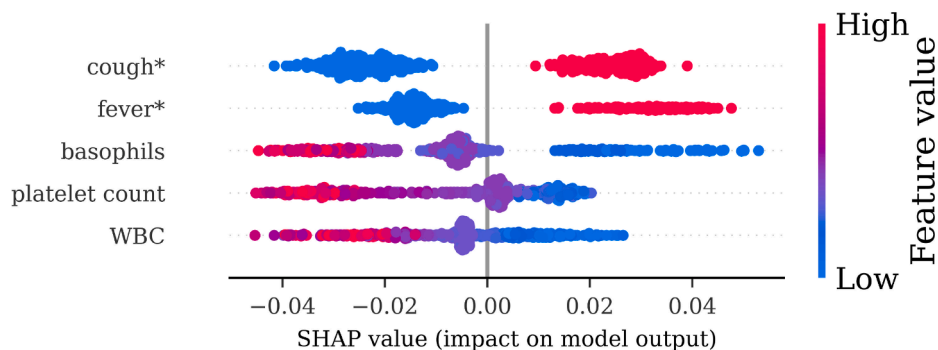


Fig. 8. SHAP plot for a single Random Forest model from the ED note experimentation with the *all* feature set, explaining the importance of features in making predictions for the withheld test set. * indicates the feature is an automatically extracted symptom.

laboratory testing in proximity to ED visits. In ED note experimentation, over 99% of samples include vital signs and 72% include blood work. In progress and telephone note experimentation, 23-38% of samples includes vital signs and 19-26% include blood work. The automatically extracted symptoms are especially important in clinical contexts, like outpatient and tele-visit, where vital signs, laboratory results, and other structured data are less available.

Fig. 8 presents a SHAP value plot for the five most predictive features from a single Random Forest model from the ED note experimentation with the *all* feature set. In this SHAP plot, each point represents a single test prediction, and the SHAP value (x-axis) describes the feature importance. Positive SHAP values indicate support for COVID-19 positivity, and negative values indicate support for negative test result. The color coding indicates the feature value, where red indicates higher feature values and blue indicates lower feature values. For example, high and moderate basophils values (coded in red and purple, respectively) have negative SHAP values, indicating support COVID-19 negativity. Low basophils values (coded in blue) have positive SHAP values, indicating support COVID-19 positivity.

Given the relatively small sample size and low proportion of positive COVID-19 tests, the SHAP impact values presented in Fig. 8 were aggregated across repeated hold-out runs. Fig. 9 presents the averaged SHAP values for each repeated hold-out run for the eight most predictive

features for the *all* feature set. For each repeated hold-out run, the absolute value of the SHAP values were averaged, yielding a single feature score per repetition. The mean SHAP values (x-axis) represents the importance of the feature in predicting COVID-19, where positive values indicate a positive correlation between the feature values and COVID-19 positivity and negative values indicate a negative correlation. The most predictive features vary by note type, although fever is a prominent indicator of COVID-19 across note types. For each note type, the top five symptoms indicating COVID-19 positivity include: ED - fever, cough, myalgia, fatigue, and flu-like symptoms; progress - fever, myalgia, respiratory symptoms, cough, and ill; and telephone - fever, cough, myalgia, fatigue, and sore throat. The differences in symptom importance by note type reflects differences in documentation in the clinical settings (e.g., emergency department, outpatient, and tele-visit).

6. Conclusions

We present CACT, a novel corpus with detailed annotations for COVID-19 diagnoses, testing, and symptoms. CACT includes 1,472 unique notes across six note types with more than 500 notes from patients with future positive COVID-19 tests. We implement the Span-based Event Extractor, which jointly extracts all annotated phenomena, including argument types and subtypes. The Span-based Event

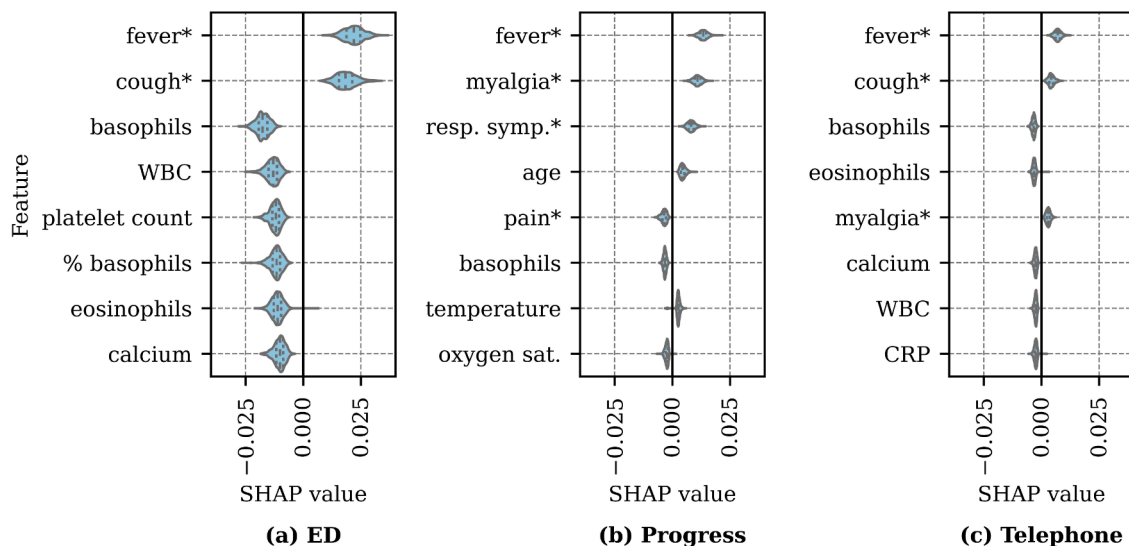


Fig. 9. Distribution of averaged SHAP values by note type with the *all* feature set. The vertical lines in each violin indicate the quartiles. * indicates the feature is an automatically extracted symptom.

Extractor achieves near-human performance in the extraction of *COVID* triggers (0.97 F1) and *Symptom* triggers (0.83 F1) and *Assertions* (0.79 F1). The performance of several attributes (e.g. *Change*, *Severity*, *Characteristics*, *Duration*, and *Frequency*) is lower than that of *Assertion*. This lower performance may partly be due to the focus on COVID-19, where clinicians' notes: (1) are highly structured around the presence/absence of a certain set of symptoms, (2) usually describe a single consultation per patient within 7 days of COVID-19 testing, and (3) focus on assessing the need for COVID-19 testing and in-person ambulatory or ED care.

In a COVID-19 prediction task, automatically extracted symptom information improved the prediction of COVID-19 test results (with significance) beyond just using structured data, and the top predictive symptoms include fever, cough, and myalgia. This application is limited by the size and scope of the available data. CACT only includes notes from early in the COVID-19 pandemic (February-March 2020), and our understanding of the presentation of COVID-19 has evolved since that time. CACT was annotated for all symptoms described in the clinical narrative, not just known symptoms of COVID-19, so the annotated symptoms cover most of the symptoms currently known to be associated with COVID-19. However, CACT includes infrequent references to losses of taste or smell. Additional annotation of notes from later in the pandemic is needed to address this gap.

In future work, the extractor will be applied to a much larger set of clinical ambulatory care and ED notes from UW. The extracted symptom information will be combined with routinely coded data (e.g. diagnosis and procedure codes, demographics) and automatically extracted data (e.g. social determinants of health [61]). Using these data, we will develop models for predicting risk of COVID-19 infection. These models could better inform clinical indications for prioritizing testing, and the presence or absence of certain symptoms can be used to inform clinical care decisions with greater precision. This future work may also identify combinations of symptoms (including their presence, absence, severity, sequence of appearance, duration, etc.) associated with clinical outcomes and health service utilization, such as deteriorating clinical course and need for repeat consultation or hospital admission. The use of detailed symptom information will be highly valuable in informing these models, but potentially only with the level of nuance that our extraction models provide. For the COVID-19 pandemic, we anticipate that the extraction model presented here will be of increasing value to clinical researchers, as the need to distinguish COVID-19 from other viral and bacterial respiratory infections becomes more necessary. As the pandemic subsides with widespread vaccination, we will return to the more typical "winter respiratory infection/influenza" seasons, where routine medical care involves differentiating COVID-19 from many other types of viral infections and identifying individuals that require COVID-19 testing. Symptom extraction models, like the model presented here, may provide the data needed to determine risk of certain infections and triage the need for testing.

We intend to explore the value of this symptom annotation scheme and extraction approach for clinical conditions where multiple consultations lead to a time point in the diagnosis pathway and symptom attributes, like change and severity, are even more important. We are especially interested in medical conditions for which delayed or missed diagnoses are known to lead to patient harm [9]. We intend to examine data sets associated with other acute and chronic conditions to investigate symptom patterns that could be used to more efficiently and accurately identify patients with these conditions. Specifically, we intend to further develop the symptom extractor to reduce diagnostic delay for lung cancer, which is known to present at a later stage. Lung cancer diagnosis often occurs after many consultations in ambulatory settings, and there may be opportunities to more quickly identify high-risk individuals based on symptoms.

Declaration of Competing Interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the National Library of Medicine Biomedical and Health Informatics Training Program under Award Number 5T15LM007442-19, the Gordon and Betty Moore Foundation, and the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1 TR002319. We want to acknowledge Elizabeth Chang, Kylie Kerker, Jolie Shen, and Erica Qiao for their contributions to the gold standard annotations and Nicholas Dobbins for data management and curation. Research and results reported in this publication was partially facilitated by the generous contribution of computational resources from the University of Washington Department of Radiology.

Appendix A

Tables 6–10.

Table 6
Expert-derived mapping of symptoms to canonical forms.

Normalized symptom	Symptom variants
altered mental status	ams, confused, confusion
anxiety	agitated, agitation, anxious
arthralgia	arthralgias
bleeding	bleed, blood, bloody
bruising	bruise, bruises, ecchymosis
chest pain	cp
chills	chill
cough	c, c., cough cough, coughing, coughs, distress coughing, distressed coughing
cramping	cramps
decreased appetite	loss of appetite, poor appetite, poor p.o. intake, poor po intake, reduced appetite
deformities	deformity
dehydration	dehydrated
diarrhea	d, d., diarrhea stools, loose stools
disharge	drainage
distended	distention
dysphagia	difficulty swallowing, dysphagia symptoms
erythema	erythematous, redness
exudates	exudate
fall	falls
fatigue	drowsiness, drowsy, fatigued, somnolence, somnolent, tired, tiredness
fever	f, f., febrile, fevers
flu-like symptoms	flu - like symptoms, influenza - like symptoms
gi symptoms	abdominal symptoms
headache	ha, headaches
heartburn	gerd symptoms, heartburn symptoms
hematochezia	brbpr
ill	ill - appearing, ill appearing, ill symptoms, illness, sick
incontinent	incontinence
irritation	irritable
itching	itchy
lethargy	lethargic
lightheadedness	dizziness, dizzy, headedness, lightheaded
myalgia	ache, aches, aching, bodyaches, myalgias
nausea	n, n., nauseated, nauseous
pain	discomfort, painful, pains
pruritus	pruritis
rash	rashes
respiratory symptoms	uri symptoms
runny nose	rhinorrhea
seizures	seizure, seizures

(continued on next page)

Table 6 (continued)

Normalized symptom	Symptom variants
shortness of breath	__shortness of breath, difficult breathing, difficulty breathing, difficulty of breathing, distress breathing, distressed breathing, doe, dsypnea, dyspnea, dyspnea, dyspnea exertion, dyspnea on exertion, increase work of breathing, increased work of breathing, out of breath, respiratory distress, short of breath, shortneses of breath, shortness breath, shortness of breaths, sob, sob on exertion, trouble breathing, work of breathing
sore throat	pharyngitis
soreness	sore
sputum	sputum production
sweats	diaphoresis, nightsweats, sweating
swelling	edema, oedema, swollen
syncope	fainting
tenderness	tender
tremors	tremor
ulcers	ulcer, ulceration, ulcerations
urinary symptoms	urinary
urination	urinating
vomiting	emesis, v, v., vomiting
weakness	weak
wheezing	wheeze, wheezes
wounds	wound

Table 7

Demographic, vital signs, and laboratory fields that are predictive of COVID-19 infection in current literature.

Parameter	Sources
age	[47,48,51,52]
alanine aminotransferase (ALT)	[49,51]
albumin	[50]
alkaline phosphatase (ALP)	[51]
aspartate aminotransferase (AST)	[47,49,51]
basophils	[51]
calcium	[47]
C-reactive protein (CRP)	[47,49,50]
D-dimer	[49]
eosinophils	[49,51]
gamma-glutamyl transferase (GGT)	[51]
gender	[52]
heart rate	[47]
lactate dehydrogenase (LDH)	[49–51]
lymphocytes	[48–52]
monocytes	[51]
neutrophils	[48,49,51,52]
oxygen saturation	[47]
platelets	[51]
prothrombin time (PT)	[49]
respiratory rate	[47]
temperature	[47,48,52]
troponin	[49]
white blood cell (WBC) count	[47,51,52]

Table 8

Hyperparameters for the Span-based Event Extractor.

Parameter	Value
Maximum sentence length, n	30
Maximum span length, M	6
Top- K spans per classifier	sentence token count
Batch size	100
Number of epochs	100
Learning rate	0.001
Optimizer	Adam
Maximum gradient L2-norm	100
BERT embedding dropout	0.3
bi-LSTM hidden size, v_h	200
bi-LSTM activation function	tanh
bi-LSTM dropout	0.3
Span classifier projection size, v_s	100
Span classifier activation function	ReLU
Span classifier dropout	0.3
Role classifier projection size, v_r	100
Role classifier activation function	ReLU
Role classifier dropout	0.3

Table 9

Structured fields from UW EHR used to predict COVID-19 infection. f indicates the function used to aggregate multiple measurements/values. Fields that measure the same phenomena and were treated as a single feature, resulting in 29 distinct structured EHR fields: {"Temperature - C," "Temperature (C)"}, {"HR," "Heart Rate"}, and {"O2 Saturation (%)," "Oxygen Saturation"}. All fields numerical (e.g. "Temperature (C)" =38.1), except "Troponin I Interpretation" and "Gender".

Parameter	Fields in UW EHR	f
age	"AgeIn2020"	max
ALT	"ALT (GPT)"	max
albumin	"Albumin"	min
ALP	"Alkaline Phosphatase (Total)"	max
AST	"AST (GOT)"	max
basophils	"Basophils" and "% Basophils"	min
calcium	"Calcium"	min
CRP	"CRP, high sensitivity"	max
D-dimer	"D_Dimer Quant"	max
eosinophils	"Eosinophils" and "% Eosinophils"	min
GGT	"Gamma Glutamyl Transferase"	max
gender	"Gender"	last
heart rate	"Heart Rate" and "HR"	max
LDH	"Lactate Dehydrogenase"	max
lymphocytes	"Lymphocytes" and "% Lymphocytes"	min
monocytes	"Monocytes"	max
neutrophils	"Neutrophils" and "% Neutrophils"	max
oxygen saturation	"Oxygen Saturation" and "O2 Saturation (%)"	min
platelets	"Platelet Count"	min
PT	"Prothrombin Time Patient" and "Prothrombin INR"	max
respiratory rate	"Respiratory Rate"	max
temperature	"Temperature - C" and "Temperature (C)"	max
troponin	"Troponin_I" and "Troponin_I Interpretation"	max
WBC count	"WBC"	min

Table 10
COVID-19 prediction hyperparameters for Random Forest Models.

Note type	Features	# estimators	Maximum depth	Minimum samples per split	Minimum samples per leaf	Class weight ratio (pos./neg.)
ED	structured	200	4	2	1	10
ED	notes	200	4	4	1	6
ED	all	200	6	3	1	8
Progress	structured	200	18	6	1	6
Progress	notes	200	10	4	1	4
Progress	all	200	8	4	1	6
Telephone	structured	200	10	2	1	2
Telephone	notes	200	6	2	1	4
Telephone	all	200	10	8	1	4

References

- [1] World Health Organization, Coronavirus disease (COVID-19) Weekly Epidemiological Update, 20 December 2020, <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>, 2020a.
- [2] H. Rossman, A. Keshet, S. Shilo, A. Gavrieli, T. Bauman, O. Cohen, E. Shelly, R. Balicer, B. Geiger, Y. Dor, et al., A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys, *Nat. Med.* (2020) 1–4, <https://doi.org/10.1038/s41591-020-0857-9>.
- [3] Z. Wu, J.M. McGoogan, Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention, *J. Am. Med. Assoc.* 323 (13) (2020) 1239–1242, <https://doi.org/10.1001/jama.2020.2648>.
- [4] J. Yang, Y. Zheng, X. Gou, K. Pu, Z. Chen, Q. Guo, R. Ji, H. Wang, Y. Wang, Y. Zhou, Prevalence of comorbidities in the novel Wuhan coronavirus (COVID-19) infection: a systematic review and meta-analysis, *International Journal of Infectious Diseases* doi:10.1016/j.ijid.2020.03.017.
- [5] P. Vetter, D.L. Vu, A.G. L'Huillier, M. Schibler, L. Kaiser, F. Jacqueroiz, Clinical features of COVID-19, *Brit. Med. J.* doi:10.1136/bmj.m1470.
- [6] G. Qian, N. Yang, A.H.Y. Ma, L. Wang, G. Li, X. Chen, X. Chen, COVID-19 Transmission Within a Family Cluster by Presymptomatic Carriers in China, *Clin. Infect. Diseases* doi:10.1093/cid/ciaa316.
- [7] W.E. Wei, Z. Li, C.J. Chiew, S.E. Yong, M.P. Toh, V.J. Lee, Presymptomatic Transmission of SARS-CoV-2–Singapore, January 23–March 16, 2020, *Morb. Mortal. Wkly Rep.* 69 (14) (2020) 411, <https://doi.org/10.15585/mmwr.mm6914e1>.
- [8] C. Wu, X. Chen, Y. Cai, X. Zhou, S. Xu, H. Huang, L. Zhang, X. Zhou, C. Du, Y. Zhang, et al., Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China, *Intern. Med.* (2019), <https://doi.org/10.1001/jamainternmed.2020.0994>.
- [9] D.E. Newman-Toker, A.C. Schaffer, C.W. Yu-Moe, N. Nassery, A.S.S. Tehrani, G. D. Clemens, Z. Wang, Y. Zhu, M. Fanai, D. Siegal, Serious misdiagnosis-related harms in malpractice claims: the “Big Three”—vascular events, infections, and cancers, *Diagnosis (Berl)* 6 (3) (2019) 227–240, <https://doi.org/10.1515/dx-2019-0019>.
- [10] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) The TRIPOD Statement, *BMJ* 131 (2) (2015) 211–219, <https://doi.org/10.1136/bmj.g7594>.
- [11] L.L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R.M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D.A. Murrack, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A.D. Wade, K. Wang, N.X.R. Wang, C. Wilhelm, B. Xie, D.M. Raymond, D.S. Weld, O. Etzioni, S. Kohlmeier, COVID-19: The COVID-19 Open Research Dataset, in: *Applied Computational Linguistics Workshop on NLP for COVID-19*, Association for Computational Linguistics, Online, <https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1>, 2020a.
- [12] World Health Organization, Global literature on coronavirus disease, URL <https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/>, 2020b.
- [13] X. Wang, X. Song, Y. Guan, B. Li, J. Han, Comprehensive named entity recognition on COVID-19 with distant or weak supervision, *arXiv* <https://arxiv.org/abs/2003.12218>.
- [14] B.R. South, S. Shen, M. Jones, J. Garvin, M.H. Samore, W.W. Chapman, A.V. Gundlapalli, Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease, *BMC Bioinform.*, 10, doi:10.1186/1471-2105-10-s9-s12.
- [15] R. Koeling, J. Carroll, R. Tate, A. Nicholson, Annotating a corpus of clinical text records for learning to recognize symptoms automatically, in: *International Workshop on Health Text Mining and Information Analysis*, 43–50, <http://sro.sussex.ac.uk/id/eprint/22351>, 2011.
- [16] Ö. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, *J. Am. Med. Inform. Assoc.* 18 (5) (2011) 552–556, <https://doi.org/10.1136/amiajnl-2011-000203>.
- [17] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C. G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, *J. Am. Med. Inform. Assoc.* 17 (5) (2010) 507–513, <https://doi.org/10.1136/jamia.2009.001560>.
- [18] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, in: *American Medical Informatics Association Annual Symposium*, 17, URL <https://www.ncbi.nlm.nih.gov/pubmed/11825149>, 2001.
- [19] D. Demner-Fushman, W.J. Rogers, A.R. Aronson, MetaMap Lite: an evaluation of a new Java implementation of MetaMap, *J. Am. Med. Inform. Assoc.* 24 (4) (2017) 841–844, <https://doi.org/10.1093/jamia/ocw177>.
- [20] M. Yetisgen, L. Vanderwende, T. Black, S. Mooney, P. Tarczy-Hornoch, A New Way of Representing Clinical Reports for Rapid Phenotyping, in: *AMIA Joint Summits on Translational Science*, 2016.
- [21] C.A. Bejan, L. Vanderwende, F. Xia, M. Yetisgen-Yildiz, Assertion modeling and its role in clinical phenotype identification, *J. Biomed. Inform.* 46 (1) (2013) 68–74, <https://doi.org/10.1016/j.jbi.2012.09.001>.
- [22] A. Wen, S. Fu, S. Moon, M. El Wazir, A. Rosenbaum, V.C. Kaggal, S. Liu, S. Sohn, H. Liu, J. Fan, Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation, *npj Digital Med.* 2 (1) (2019) 1–7, <https://doi.org/10.1038/s41746-019-0208-8>.
- [23] Mayo Clinic NLP, OHNLP/MedTagger, <https://github.com/OHNLP/MedTagger>, 2020.
- [24] S. Zheng, Y. Hao, D. Lu, H. Bao, J. Xu, H. Hao, B. Xu, Joint entity and relation extraction based on a hybrid neural network, *Neurocomputing* 257 (2017) 59–66, doi:10.1016/j.neucom.2016.12.075, machine Learning and Signal Processing for Big Multimedia Analysis.
- [25] W. Orr, P. Tadepalli, X. Fern, Event Detection with Neural Networks: A Rigorous Empirical Evaluation, in: *Conference on Empirical Methods in Natural Language Processing*, 999–1004, doi:<https://doi.org/10.18653/v1/D18-1122>, 2018.
- [26] X. Shi, Y. Yi, Y. Xiong, B. Tang, Q. Chen, X. Wang, Z. Ji, Y. Zhang, H. Xu, Extracting entities with attributes in clinical text via joint deep learning, *J. Am. Med. Inform. Assoc.* 26 (12) (2019) 1584–1591, <https://doi.org/10.1093/jamia/ocz158>.
- [27] Y. Pang, J. Liu, L. Liu, Z. Yu, K. Zhang, A Deep Neural Network Model for Joint Entity and Relation Extraction, *IEEE Access* 7 (2019) 179143–179150, <https://doi.org/10.1109/ACCESS.2019.2949086>.
- [28] L. Chen, Y. Gu, X. Ji, Z. Sun, H. Li, Y. Gao, Y. Huang, Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning, *J. Am. Med. Inform. Assoc.* 27 (1) (2019) 56–64, <https://doi.org/10.1093/jamia/ocz141>.
- [29] F. Christopoulou, T.T. Tran, S.K. Sahu, M. Miwa, S. Ananiadou, Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods, *J. Am. Med. Inform. Assoc.* 27 (1) (2020) 39–46, <https://doi.org/10.1093/jamia/ocz101>.
- [30] K. Lee, L. He, M. Lewis, L. Zettlemoyer, End-to-end Neural Coreference Resolution, in: *Empirical Methods in Natural Language Processing*, 188–197, 2017, doi:10.18653/v1/D17-1018.
- [31] Y. Luan, L. He, M. Ostendorf, H. Hajishirzi, Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction, in: *Empirical Methods in Natural Language Processing*, 3219–3232, 2018, doi:10.18653/v1/D18-1360.
- [32] Y. Luan, D. Wadden, L. He, A. Shah, M. Ostendorf, H. Hajishirzi, A general framework for information extraction using dynamic span graphs, in: *North American Chapter of the Association for Computational Linguistics*, 3036–3046, 2019, doi:10.18653/v1/N19-1308.
- [33] D. Wadden, U. Wennberg, Y. Luan, H. Hajishirzi, Entity, Relation, and Event Extraction with Contextualized Span Representations, in: *Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, 5788–5793, 2019, doi:10.18653/v1/D19-1585.
- [34] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: *North American Chapter of the Association for Computational Linguistics*, 2227–2237, 2018, doi:10.18653/v1/N18-1202.
- [35] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *North American Chapter of the Association for Computational Linguistics*, 4171–4186, 2019, doi:10.18653/v1/N19-1423.
- [36] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, Q.V. Le, XLNet: Generalized autoregressive pretraining for language understanding, in: *Advances in neural information processing systems*, 5753–5763, <http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf>, 2019.

- [37] W. Huang, X. Cheng, T. Wang, W. Chu, BERT-Based Multi-Head Selection for Joint Entity-Relation Extraction, in: International Conference on Natural Language Processing and Chinese Computing, 713–723, 2019, doi:10.1007/978-3-030-32236-6_65.
- [38] H. Wang, M. Tan, M. Yu, S. Chang, D. Wang, K. Xu, X. Guo, S. Potdar, Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers, in: Association for Computational Linguistics, 1371–1377, doi:10.18653/v1/P19-1132, 2019.
- [39] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly Available Clinical BERT Embeddings, in: Clinical Natural Language Processing Workshop, 72–78, 2019, doi:10.18653/v1/W19-1909.
- [40] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, *Scientific Data* 3 (2016) 160035, <https://doi.org/10.1038/sdata.2016.35>.
- [41] W. Tian, W. Jiang, J. Yao, C.J. Nicholson, R.H. Li, H.H. Sigurslid, L. Wooster, J.I. Rotter, X. Guo, R. Malhotra, Predictors of mortality in hospitalized COVID-19 patients: A systematic review and meta-analysis, *Journal of Medical Virology* doi: 10.1002/jmv.26050.
- [42] S. Figliozzi, P.G. Masci, N. Ahmadi, L. Tondi, E. Koutli, A. Aimò, K. Stamatelopoulos, M.-A. Dimopoulos, A.L. Caforio, G. Georgiopoulos, Predictors of adverse prognosis in COVID-19: A systematic review and meta-analysis, *Eur. J. Clin. Invest.* (2020) e13362, <https://doi.org/10.1111/eci.13362>.
- [43] V. Jain, J.-M. Yuan, Predictive symptoms and comorbidities for severe COVID-19 and intensive care unit admission: a systematic review and meta-analysis, *Int. J. Public Health* (2020) 1, <https://doi.org/10.1007/s00038-020-01390-7>.
- [44] Y. Dong, H. Zhou, M. Li, Z. Zhang, W. Guo, T. Yu, Y. Gui, Q. Wang, L. Zhao, S. Luo, et al., A novel simple scoring model for predicting severity of patients with SARS-CoV-2 infection, *Transboundary and Emerging Diseases* doi:10.1111/tbed.13651.
- [45] P.P. Xu, R.H. Tian, S. Luo, Z.Y. Zu, B. Fan, X.M. Wang, K. Xu, J.T. Wang, J. Zhu, J. C. Shi, et al., Risk factors for adverse clinical outcomes with COVID-19 in China: a multicenter, retrospective, observational study, *Theranostics* 10 (14) (2020) 6372, <https://doi.org/10.7150/thno.46833>.
- [46] J.L. Izquierdo, J. Ancochea, J.B. Soriano, S.C.-. R. Group, et al., Clinical Characteristics and Prognostic Factors for Intensive Care Unit Admission of Patients With COVID-19: Retrospective Study Using Machine Learning and Natural Language Processing, *J. Med. Internet Res.* 22 (10) (2020) e21801, doi:10.2196/21801.
- [47] D. Bertsimas, L. Boussioux, R.C. Wright, A. Delarue, V.D. Jr., A. Jacquillat, D.L. Kitane, G. Lukin, M.L. Li, L. Mingardi, O. Nohadani, A. Orfanoudaki, T. Papalexopoulos, I. Paskov, J. Pauphilet, O.S. Lami, B. Stellato, H.T. Bouvardi, K.V. Carballo, H. Wiberg, C. Zeng, From predictions to prescriptions: A data-driven response to COVID-19, arXiv preprint 2006.16509, <https://arxiv.org/abs/2006.16509>.
- [48] L. Wynants, B. Van Calster, G.S. Collins, R.D. Riley, G. Heinze, E. Schuit, M.M.J. Bonten, J.A.A. Damen, T.P.A. Debray, M. De Vos, P. Dhiman, M.C. Haller, M.O. Harhay, L. Henckaerts, N. Kreuzberger, A. Lohmann, K. Luijken, J. Ma, C.L. Andaur Navarro, J.B. Reitsma, J.C. Sergeant, C. Shi, N. Skoetz, L.J.M. Smits, K.I.E. Snell, M. Sperrin, R. Spijker, E.W. Steyerberg, T. Takada, S.M.J. van Kuijk, F.S. van Royen, C. Wallisch, L. Hoof, K.G.M. Moons, M. van Smeden, Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal, *BMJ* 369, doi: 10.1136/bmj.m1328.
- [49] J.A. Siordia, Epidemiology and clinical features of COVID-19: A review of current literature, *J. Clin. Virol.* 127 (2020) 104357, ISSN 1386-6532, doi:10.1016/j.jcv.2020.104357.
- [50] J.J. Zhang, K.S. Lee, L.W. Ang, Y.S. Leo, B.E. Young, Risk factors of severe disease and efficacy of treatment in patients infected with COVID-19: A systematic review, meta-analysis and meta-regression analysis, *Clin. Infect. Diseases* doi:10.1093/cid/ciaa576.
- [51] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, F. Cabitza, Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study, *J. Med. Syst.* 44 (8) (2020) 135, <https://doi.org/10.1007/s10916-020-01597-4>.
- [52] X. Mei, H.-C. Lee, K.-Y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, Y. Ma, P.M. Robson, M. Chung, A. Bernheim, V. Mani, C. Calcagno, K. Li, S. Li, H. Shan, J. Lv, T. Zhao, J. Xia, Q. Long, S. Steinberger, A. Jacobi, T. Deyer, M. Luksza, F. Liu, B.P. Little, Z. A. Fayad, Y. Yang, Artificial intelligence-enabled rapid diagnosis of patients with COVID-19, *Nat. Med.* 26 (8) (2020) 1224–1228, <https://doi.org/10.1038/s41591-020-0931-3>.
- [53] S. Wollenstein-Betech, C. Cassandras, I. Paschalidis, Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: Hospitalizations, mortality, and the need for an ICU or ventilator., *Int. J. Med. Inform.* doi:10.1016/j.ijmedinf.2020.104258.
- [54] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, BRAT: A Web-based Tool for NLP-Assisted Text Annotation, in: Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107.
- [55] C. Walker, S. Strassel, J. Medero, K. Maeda, ACE 2005 Multilingual Training Corpus LDC2006T06, <https://catalog.ldc.upenn.edu/LDC2006T06>, 2006.
- [56] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), *Advances in Neural Information Processing Systems* 32, Curran Associates Inc, 8024–8035, <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>, 2019.
- [57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine Learning in Python, *J. Mach. Learn. Res.* 12 (2011) 2825–2830, <https://doi.org/10.5555/1953048.2078195>.
- [58] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, *Nat. Mach. Intell.* 2 (1) (2020) 2522–2539, <https://doi.org/10.1038/s42256-019-0138-9>.
- [59] J.-H. Kim, Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap, *Comput. Stat. Data Anal.* 53 (11) (2009) 3735–3745, <https://doi.org/10.1016/j.csda.2009.04.009>.
- [60] D.S. Smith, E.A. Richey, W.L. Brunetto, A Symptom-Based Rule for Diagnosis of COVID-19, *SN Comprhen. Clin. Med.* 2 (11) (2020) 1947–1954, <https://doi.org/10.1007/s42399-020-00603-7>.
- [61] K. Lybarger, M. Ostendorf, M. Yetisgen, Annotating Social Determinants of Health Using Active Learning, and Characterizing Determinants Using Neural Event Extraction, *J. Biomed. Inform.* doi:10.1016/j.jbi.2020.103631.