Learning Clinical Concepts for
Predicting Risk of Progression to Severe COVID-19

Helen Zhou, M.S., M.Eng.,[1] Cheng Cheng, B.S.,[1] Kelly J. Shields, Ph.D.,[2]
Gursimran Kochhar M.D., F.A.C.P., CN.S.C.,[3] Tariq Cheema, M.D., F.C.C.P., M.M.M.,[3]
Zachary C. Lipton, Ph.D.,[1] Jeremy C. Weiss, M.D., Ph.D.[1]
[1]Carnegie Mellon University, Pittsburgh, PA; [2]Highmark Health Enterprise Data &
Analytics, Data Science R&D; [3]Allegheny Health Network, Pittsburgh, PA

Abstract

With COVID-19 now pervasive, identification of high-risk individuals is crucial. Using data from a major healthcare provider in Southwestern Pennsylvania, we develop survival models predicting severe COVID-19 progression. In this endeavor, we face a tradeoff between more accurate models relying on many features and less accurate models relying on a few features aligned with clinician intuition. Complicating matters, many EHR features tend to be under-coded degrading the accuracy of smaller models. In this study we develop two sets of high-performance risk scores: (i) an unconstrained model built from all available features; and (ii) a pipeline that learns a small set of clinical concepts before training a risk predictor. Learned concepts boost performance over the corresponding features (C-index 0.858 vs. 0.844) and demonstrate improvements over (i) when evaluated out-of-sample (subsequent time periods). Our models outperform previous works (C-index 0.844–0.872 vs. 0.598–0.810).

Introduction

As COVID-19 becomes endemic, communities are learning what it means for them to "live with" COVID-19. An important component of living with COVID-19 is understanding when individuals who contract the disease are likely to progress to a severe condition. Our work studies the risk of severe COVID-19 progression, using data collected by a major healthcare provider in Southwestern Pennsylvania from January 2020 to January 2022. We define *severe COVID-19* as a COVID-19 case involving mechanical ventilation, admission to an intensive care unit (ICU), or death.

Healthcare providers often have different systems for collecting and storing patient data. To utilize this data for risk estimation, researchers usually leverage domain expertise to manually extract a large initial set of potentially relevant features, and subsequently use automatic feature selection techniques to eliminate all but the most significant features. Clinicians can then consider these features when determining a patient's care plan, and hospitals could potentially extract these features in order to calculate risk.

While expert-guided curation of features can help reduce the model search space, it can also limit performance due to imperfect feature extraction and inadvertent removal of informative features. Automatic feature selection may yield features that are predictive of the outcome, but these features may actually be serving as *proxies* for higher-level concepts that cause the outcome (e.g. insulin medication may be predictive, but diabetes is the underlying risk factor), *especially when the higher-level concepts are underreported*. While reliably recorded proxies can be effective predictors, they can also yield misleading interpretations. Additionally, it is unclear whether these proxies will be equally effective when applied to new settings such as different time periods or different hospitals. As a result, doctors may favor smaller models with features whose relevance is intuitive even if these models suffer some loss in performance owing, in part, to underreporting of the features.

To strike a balance between incorporating domain knowledge, model simplicity, transparency, and performance, we propose to learn clinical concepts anchored to intuitive expert-selected features, and to use these concepts to predict severe COVID-19 progression. Motivated by high levels of missingness in our data, clinical concepts are learned by treating the presence of an expert-selected feature (e.g. diabetes ICD code) as a positive label, treating its absence as unlabeled, and applying positive and unlabeled learning algorithms to learn the probability of the concept given the other covariates. We find that learned concepts (LC) for an expert-selected subset of features provide a boost in performance over the features (C-index 0.858 vs. 0.844), and that this boost places the LC model approximately halfway between the selected features model (C-index 0.844) and the model trained on all available features (All Features) or LC + All Features combined (both C-index 0.872). While there is some loss of performance going from the All Features model to the LC model, this gap seems to close quickly on subsequent time periods, suggesting that there may be some reason why the LC model is favored over time. Qualitatively, we find that some of the features important to the All Features model are incorporated into the learned concept classifiers, possibly indicating that they serve as proxies for the concepts. Finally, we publish an interactive web visualization tool at acmilab.org/severe_covid for users to explore the learned concepts, original features, and how both are utilized in our models.

Related Work

Several works have identified predictive factors for severe COVID-19, where the population studied and the definition of severity vary. Docherty et. al.[1] performed a prospective observational study on COVID-19 hospitalizations in the UK and identified risk factors for mortality including old age, male, and chronic comorbidities such as obesity. Henry et. al.[2] performed a meta-analysis of 21 studies and identified white blood cell count, lymphocytes, platelets, IL-6 and serum ferritin as inpatient biomarkers for progression to severe or fatal illness. The VACO Index[3] uses three pre-COVID-19 health status variables, demographics, pre-existing medical conditions, and Charlson Comorbidity Index, in a mortality score. To identify severe COVID-19 patients in need of limited ventilation resources, some works[4,5] have predicted patient risk of developing acute respiratory distress syndrome (ARDS) using labs, demographics, and other clinical data. Covichem[6] is an admission risk score predicting severity as defined by a composite of lab values, ARDS, or ICU admission. After stepwise model selection on the Akaike Information Criterion, Covichem[6] identified risk factors including: obesity, cardiovascular conditions, plasma sodium, albumin, ferritin, lactate, and creatinine. COVID-GRAM[7] predicts risk of ICU admission, invasive ventilation, or mortality for inpatients using ten predictors: chest radiography abnormality, age, hemoptysis, dyspnea, unconsciousness, comorbidity count, cancer history, neutrophil-to-lymphocyte ratio, lactate dehydrogenase, and direct bilirubin, chosen via LASSO regression. Galloway et. al.[8] created a simple count-based risk score for predicting ICU admission or mortality, using twelve features: age, male, ethnicity, oxygen saturation, radiological severity score, neutrophils, C-reactive protein, albumin, creatinine, diabetes mellitus, hypertension, and chronic lung disease. Other works[9,10] have used chest CTs to score severity.

Several works have used deep learning to extract embeddings of medical concepts from EHRs.[11,12] While useful for various downstream tasks, these embeddings usually suffer a lack of transparency. As an alternative, Halpern et. al. proposed an "anchor-and-learn" framework in which expert-defined binary medical concepts are learned by treating certain informative features as positive labels for those concepts, and applying algorithms from positive and unlabeled learning.[13–15] An advantage of this method is the interpretable coefficients of classifiers used for learning the concepts.

Data

Cohort Description. We use retrospective observational data collected by a major healthcare provider in Southwestern Pennsylvania from January 1st, 2020 to January 12th, 2022. Out of 171,009 patients who were tested for COVID-19, we extract the 40,190 who tested positive. Of those, we remove individuals who were already mechanically ventilated or admitted to the ICU within 30 days prior to the time t_0 of testing positive for the first time. This leaves a cohort of 31,336 individuals (Table 1). Note that this study seeks to predict the risk of progressing to severe COVID-19 *upon testing positive for the first time*, and so features and outcomes are defined relative to time t_0.

Features. Features are extracted no later than the date of each patient's first covid positive test. These include testing location (inpatient/outpatient), demographics, labs, medications, vaccines, symptoms, and problem history. The most recent value of each feature is extracted, and symptoms are limited to a one-day window around t_0. Since there are tens of thousands of distinct medications, labs, diagnoses, vaccines, etc. in our data, the feature pool is limited to the top 20 of each data type except for labs (top 50). Upon clinician review, 45 more features are extracted. After removing low-variance features, converting categorical values to indicators, and normalizing continuous values, this yields a fixed-length 139-dimensional feature vector (see acmilab.org/severe_covid) for patient information known at t_0.

Outcome. Since patients are right-censored upon leaving the hospital system, the outcome of interest is a time-to-event, where the time is computed as the time elapsed between t_0 and severe COVID-19 (mechanical ventilation, ICU admission, or death) or censorship (when the patient was last seen in hospital records), whichever is first.

Table 1. Cohort characteristics (n = 31,336). Demographics, inpatient vs. outpatient status at t_0, and outcomes.

Characteristic		Count (%)	Characteristic (cont.)		Count (%)
Gender	Female	17,874 (57.0%)	*Location of test*	Inpatient	13,246 (42.3%)
	Male	13,455 (42.9%)		Outpatient	15,868 (50.6%)
Age	Under 20	2,836 (9.1%)		Unknown	2,222 (7.1%)
	20 – 30	3,987 (12.7%)	*Outcomes*	Severe COVID-19	5,272 (16.8%)
	30 – 40	4,134 (13.2%)		ICU Admission	4,811 (15.4%)
	40 – 50	4,155 (13.3%)		Death	1,554 (5.0%)
	50 – 60	5,444 (17.4%)		Mechanical ventilation	1,096 (3.5%)
	60 – 70	5,017 (16.0%)			
	70 or above	5,763 (18.4%)			

Learning Clinical Concepts

Different types of data often provide partial information about higher-level concepts. For example, a saline IV bolus is typically administered inpatient, and is highly predictive of inpatient status even if inpatient status is unavailable. Certain labs could further confirm inpatient status. While one could methodically create rules for every concept of interest, it is difficult to do so comprehensively. As a result, learned models may end up using proxies that indirectly encode important risk factors (e.g. IV bolus encoding inpatient status), possibly leading to misinterpretation. Thus, we learn clinical concepts corresponding to major risk factors, and use these for downstream risk prediction.

PU Algorithm for Learning Concepts. To learn these concepts, we use the "anchor-and-learn" framework.[14] For each concept of interest, we identify some key informative observations ("positive anchors") relating to that concept. In this work, we only consider binary-valued concepts (present vs. not present). An observation is an *anchor* for a concept if it is conditionally independent of all other observations conditioned on the concept. When the presence of an anchor almost certainly implies the presence of the concept, this is known as a *positive anchor*.

Consider a patient with covariates $x \in R^d$. Suppose we want to extract a concept c with positive anchor $x_c \in \{0,1\}$ (e.g. extracting a diabetes concept with a diabetes diagnosis code as a positive anchor). Let $y_c \in \{0,1\}$ be the true binary label for whether concept c is present. Note that in most observational health data, we observe the presence of a clinical condition, but not the absence of it. For example, when extracting the diabetes concept, we can be fairly confident that a patient marked as diabetic does indeed have diabetes, but patients unmarked do not necessarily *not have* diabetes. Said differently, we have positive and unlabeled (PU) data rather than positive and negative data. Since only positive examples are labeled, $y_c = 1$ is certain when $x_c = 1$, but when $x_c = 0$, then y_c could be either 0 or 1.

Thus, we leverage algorithms designed to learn from positive and unlabeled data, or "PU learning" algorithms. Let $x_{\bar{c}}$ refer to all covariates except for x_c. Since anchors are conditionally independent of all other observations conditioned on the concept, we have that $p(x_c|y_c = 1) = p(x_c|y_c = 1, x_{\bar{c}})$. Now, consider $p(x_c = 1|x_{\bar{c}})$. We have that:

$$p(x_c = 1|x_{\bar{c}}) = p(x_c = 1 \wedge y_c = 1| x_{\bar{c}})$$
$$= p(y_c = 1|x_{\bar{c}})p(x_c = 1|y_c = 1, x_{\bar{c}})$$
$$= p(y_c = 1|x_{\bar{c}})p(x_c = 1|y_c = 1)$$
$$\Rightarrow \boldsymbol{p(y_c = 1|x_{\bar{c}}) = p(x_c = 1|x_{\bar{c}})/\delta_c}$$

where $\delta_c = p(x_c = 1|y_c = 1)$. The first equality follows from the fact that $y_c = 1$ is certain when $x_c = 1$, and the second equality follows from Bayes rule. In words, the expression indicates that *true probability of the concept* being present is *proportional to the probability of the positive anchor being present* by a factor of δ_c. Thus, if we can train a PU classifier $g(x_{\bar{c}}) = p(x_c = 1|x_{\bar{c}})$ that learns the probability that a positive anchor is present given the remaining covariates, we need only scale the probability by δ_c in order to get the probability of the underlying concept being present. As noted in Elkan and Noto,[13] for the set P of positive labeled examples, one can construct an empirical estimate of the constant δ_c as $\hat{\delta}_c = \frac{1}{n}\sum_{x_{\bar{c}}\in P} g(x_{\bar{c}})$, due to the observation that $g(x_{\bar{c}}) = \delta_c$ for $x_{\bar{c}} \in P$:

$$g(x_{\bar{c}}) = p(x_c = 1|x_{\bar{c}})$$
$$= p(x_c = 1|x_{\bar{c}}, y_c = 1)p(y_c = 1|x_{\bar{c}}) + p(x_c = 1|x_{\bar{c}}, y_c = 0)p(y_c = 0|x_{\bar{c}})$$
$$= p(x_c = 1|x_{\bar{c}}, y_c = 1) \cdot 1 + 0 \cdot 0 \quad \text{since } x_{\bar{c}} \in P$$
$$= p(x_c = 1|y_c = 1)$$
$$= \delta_c.$$

Finally, this yields the following procedure for learning clinical concepts:

1. **Identify clinical concepts** of interest, and corresponding positive anchors.
2. **Learn a positive vs. unlabeled classifier.** Use logistic regression to learn a classifier $g(x_{\bar{c}})$ that outputs the probability of the positive anchor given the other covariates.
3. **Estimate the scaling constant.** On a validation set, estimate $\hat{\delta}_c$ by averaging the output of $g(x_{\bar{c}})$ on all positive labeled examples (i.e. examples with the positive anchor).
4. **Scale predictions from the PU classifier** by the estimated scaling constant to get the probability that the underlying concept is present. That is, compute $p(y_c = 1|x_{\bar{c}}) = g(x_{\bar{c}})/\hat{\delta}_c$ for all examples where the positive anchor is not present. Where the positive anchor is present, leave the probability as 1.

This procedure is also used in Halpern et. al.,[14] except instead of drawing the concept from a Bernoulli distribution parameterized by $p(y_c = 1|x_{\bar{c}})$, we directly use the computed probability $p(y_c = 1|x_{\bar{c}})$ since it can provide more granular information. The *scikit-learn*[16] python package was utilized for its logistic regression implementation.

Identifying Concepts of Interest.

In order to define clinical concepts of interest, we surveyed several clinicians in the healthcare provider network about the main concepts they would look for when assessing risk of severe COVID-19. The survey yielded 21 concepts: old age, inpatient, outpatient, diabetes, shortness of breath, fever, cough, fatigue, COVID-19 vaccination, flu vaccination, obesity, hypertension, immunocompromised, COPD, congestive heart failure, chronic kidney disease, hyperglycemia, transplant, cancer, lung disease, and myalgia. We identify positive anchors for these concepts (precise definitions at acmilab.org/severe_covid) and apply the PU algorithm to extract a more complete representation of the concepts.

Learning Severe COVID-19 Risk

Using the *lifelines* python package,[17] a Cox proportional hazards model with L1 regularization (*Lasso-Cox*) is used to model risk of progression to severe COVID-19. For a patient with covariates X, their hazard h at time t is given by:

$$h(t) = h_0(t) \exp(X\beta)$$

where h_0 is a baseline hazard function, and β are learned coefficients. The regularization penalty is given by $\lambda ||\beta||_1$, where regularization strength λ is selected using 5-fold cross validation and grid search over penalties between 0 and 0.2, with a step size of 0.001 between each penalty. For stability of training, features with variance < 0.01 are removed.

Experimental Setup

Feature Sets. In order to explore the marginal effect of incorporating learned concepts versus the original set of 139 features, we analyze and evaluate Lasso-Cox models learned from five different sets of features:

(1) **Raw positive anchors**: only the positive anchors identified in the data, without learning the corresponding clinical concepts (e.g. "diabetes" being mentioned in a note, the ICD code for diabetes, etc.)
(2) **Learned concepts (LC)**: only the learned clinical concepts (e.g. the diabetes concept)
(3) **LC + Numeric**: the learned clinical concepts, as well as numerical features (e.g. diabetes concept and labs)
(4) **LC + All Features**: the learned concepts, as well as all of the original 139 features
(5) **All Features**: all 139 original features, no learned concepts.

Back-testing and Data Splits. In real-world settings, hospital systems may want to use updated data to revise their models. To emulate this process, we re-train models (including PU concept classifiers) up to the end of each 3-month season (spring, summer, fall, winter), and evaluate their performance on subsequent seasons. Spring is March 20th until June 21st, followed by summer until September 22nd, followed by fall until December 21st, followed by winter until March 20th of the following year. For each 3-month period, a 70-30 split designates train and test sets, where test data is never included in any model training. To keep the risk score interpretation simple, for each time period a grid search on the Lasso-Cox penalty is done to choose a model with approximately ten features. We additionally train models on the entire study time range, with train and test sets that aggregate the respective 3-month datasets.

Evaluation

Clinical Concept Evaluation. Since the concepts are only positively labeled or unlabeled, it is not possible to compute precision of the concept classifiers.[15] However, we can compute recall as the proportion of known positives recovered by the classifiers. Additionally, we examine the number of previously unlabeled samples predicted to be positive.

Model Interpretation. The Lasso-Cox model coefficients are in terms of original features as well as learned concepts. In addition to listing the Lasso-Cox coefficients, we create an interactive Sankey diagram to visualize how raw features translate into concepts, and how the resulting models pull from both. This gives the user a birds-eye view of how each concept is defined, the strength and sign of the coefficients, and which concepts are used in different models.

Survival Model Evaluation Metrics. The concordance, or C-index, is used to evaluate the model's discriminative ability. To evaluate calibration, both one-calibration at 14 days and D-calibration are used.[18] Additionally, low, medium, and high-risk strata are defined and their 14-day Kaplan-Meier survival curves are inspected.

Baselines. We compare our model performance to that of the Covichem[6] and Galloway[8] risk scores. For Covichem, in order to conduct a fairer comparison than directly applying their logistic regression coefficients learned on a different population, we extract the same features and re-train logistic regression on our own training data. For Galloway, we extract all but one of their twelve features (radiological severity is not available in our data), and compare performance against two versions of their model: (1) directly applying their proposed count-based risk score (Galloway count), and (2) re-training a logistic regression model using the twelve variables (Galloway reweighted).

Results

The PU learning algorithm yields concepts ranging from those with high recall, e.g. 0.974 for inpatient status, to low recall, e.g., 0.381 for immunocompromised (Table 2). Some concepts have substantially more new positives (obesity, with 2,157 new positives). Concept classifier coefficients are available at acmilab.org/severe_covid.

The model trained on learned concepts (LC) achieves a higher aggregate concordance than the original features corresponding to those concepts (0.858 vs. 0.844, Table 3). The model learned from all original features (All Features) and LCs (C-index 0.872) performs comparably to All Features alone (C-index 0.872). The addition of numerical features to the LCs does not significantly improve performance (C-index of both are 0.858). In all models except for the ones trained on All Features or All Features + LCs, the aggregate C-index is higher than the C-indices on the inpatient and outpatient subpopulations.

In the models using all features, medications such as dexamethasone, acetaminophen, and intravenous saline are selected (Table 4). Across all models, the inpatient status is the feature with the greatest hazard ratio. Blood urea nitrogen is used by both the LC + All Features and All Features models. Figure 1 is a screenshot of an interactive Sankey diagram which allows users to explore the coefficients for both the underlying clinical concepts and the classifiers built on top of the features and concepts. The interactive web tool is available at acmilab.org/severe_covid.

When evaluated over time, the models with learned concepts achieve higher concordance several months after the model was initially trained, whereas the All Features model achieves higher concordance in the immediate term. For example, in Spring 2020, the concordance of the All Features model trained up until the end of Spring 2020 is 0.842, compared to 0.797 in the LC only model. By Fall and Winter 2021, however, the All Features model degrades to 0.808 or stays around 0.845, whereas the LC only model actually increases concordance to 0.83 and 0.904. Reading the table from left to right, the performance of any model fluctuates no more than 0.121 over all seasons. Reading the table from top to bottom, several columns shown an increase in performance as models are trained on more recent data.

The Kaplan-Meier curves corresponding to the low, medium, and high risk groups derived from the LC + All Features model and LC only model predictions are shown in Figure 2. There is a clear separation between the survival trajectories of the different risk groups. The LC + all features model appears to slightly under-estimate the risk of the high-risk groups, whereas the LC only model appears to be better calibrated at 14 days (Figure 3). The LC only model also appears to have better d-calibration across all time points (Figure 4).

Table 2. The number of new positives extracted by PU learning in the test set, the number of clinical concepts originally in the test set (determined solely by the presence of positive anchors), and the recall of the PU classifier among known positives in the test set. Concepts with prevalence below 1.5% are omitted.

Learned Concept	New Positives (% of Test Set)	Original Positives (% of Test Set)	Recall Among Original Positives (count)
Old age	147 (1.6%)	3,158 (33.7%)	0.956 (3,019)
Inpatient	227 (2.4%)	3,914 (41.8%)	0.974 (3,813)
Outpatient	207 (2.2%)	4,811 (51.3%)	0.961 (4,623)
Diabetes	594 (6.3%)	834 (8.9%)	0.553 (461)
Fever	4,171 (44.5%)	1,021 (10.9%)	0.826 (843)
Shortness of breath	1,518 (16.2%)	1,005 (10.7%)	0.767 (771)
COVID-19 vaccination	2,128 (22.7%)	1,884 (20.1%)	0.739 (1,392)
Flu vaccine	2,326 (24.8%)	4,191 (44.7%)	0.864 (3,619)
Obesity	2,157 (23.0%)	433 (4.6%)	0.610 (264)
Immunocompromised	909 (9.7%)	168 (1.8%)	0.381 (64)
COPD	575 (6.1%)	220 (2.3%)	0.623 (137)
Hyperglycemia	496 (5.3%)	171 (1.8%)	0.737 (126)
Cough	4,023 (42.9%)	1,862 (19.9%)	0.815 (1,517)
Fatigue	2,947 (31.4%)	602 (6.4%)	0.694 (418)

Table 3. Performance of our models and baselines. The median C-index and 95% confidence intervals are reported from bootstrapping the test set with 1000 replicates. **Bold** highlights the two models with highest C-index.

Model	Aggregate Test C-index	Inpatient Test C-index	Outpatient Test C-index
Covichem	0.598 (0.580 – 0.616)	0.584 (0.569 – 0.600)	0.546 (0.509 – 0.581)
Galloway count	0.745 (0.734 – 0.757)	0.647 (0.633 – 0.662)	0.714 (0.677 – 0.750)
Galloway reweighted	0.810 (0.803 – 0.824)	0.699 (0.673–0.703)	0.764 (0.728–0.709)
Raw positive anchors	0.844 (0.836 – 0.851)	0.665 (0.650 – 0.680)	0.756 (0.709 – 0.796)
Learned concepts (LC) only	0.858 (0.851 – 0.865)	0.699 (0.685 – 0.713)	0.798 (0.757 – 0.834)
LC + numerical features	0.858 (0.851 – 0.865)	0.695 (0.681 – 0.71)	0.814 (0.777 – 0.849)
LC + all features	**0.872 (0.865 – 0.877)**	**0.715 (0.702 – 0.728)**	**0.879 (0.858 – 0.901)**
All features (no LC)	**0.872 (0.866 – 0.878)**	**0.717 (0.703 – 0.730)**	**0.880 (0.860 – 0.901)**

Table 4. Hazard ratios (HR) of LC + All Features, LC, and All Features models. Abbreviations: Med = Medication, loc. = location, Dex. = Dexamethasone sodium phosphate, APAP = acetaminophen, SOB = Shortness of breath, BUN = blood urea nitrogen, NEUT = neutrophils, Immunocomp. = immunocompromised, vax = vaccine, OP = outpatient.

| All Features + LCs		Learned Concepts (LC) Only		All Features Only	
Features	*HR (95% CI)*	*Features*	*HR (95% CI)*	*Features*	*HR (95% CI)*
(LC) Inpatient	2.31 (2.09 – 2.56)	(LC) Inpatient	7.23 (5.43 – 9.62)	(Test Location) Inpatient	3.62 (3.31 – 3.97)
(LC) SOB	1.72 (1.58 – 1.88)	(LC) Old age	2.54 (2.33 – 2.77)	(Med) Dex. 4mg/mL injection sol.	1.91 (1.77 – 2.06)
(Med) Dex. 4mg/mL injection sol.	1.54 (1.42 – 1.68)	(LC) SOB	2.31 (2.14 – 2.49)	(Med) APAP 325mg tablet	1.67 (1.51 – 1.85)
(Med) APAP 325 mg tablet	1.47 (1.34 – 1.62)	(LC) Diabetes	1.28 (1.16 – 1.41)	Age 70+	1.60 (1.49 – 1.72)
(LC) Old age	1.34 (1.25 – 1.44)	(LC) COPD	1.22 (1.13 – 1.32)	(Med) NaCl 0.9% IV sol.	1.35 (1.24 – 1.46)
(Med) NaCl 0.9 % IV sol.	1.33 (1.23 – 1.44)	(LC) Obesity	1.22 (1.14 – 1.3)	(OP ICD) SOB	1.10 (1.01 – 1.19)
(Lab) BUN	1.05 (1.01 – 1.1)	(LC) Immunocomp.	1.10 (1.04 – 1.16)	(Lab) BUN	1.10 (1.05 – 1.14)
		(LC) Fatigue	1.08 (1.02 – 1.15)	(Med) Pantoprazole 40 mg tablet	1.05 (0.97 – 1.12)
		(LC) Outpatient	1.06 (0.80 – 1.42)	(Lab) NEUT relative %	1.04 (0.99 – 1.09)
		(LC) Hyperglycemia	0.90 (0.81 – 1.00)	(Med) NaCl 0.9% IV Bolus	1.04 (0.96 – 1.13)
		(LC) COVID-19 vax	0.87 (0.78 – 0.97)	(Lab) Albumin	0.98 (0.93 – 1.02)
		(LC) Fever	0.82 (0.75 – 0.90)		
		(LC) Flu vax	0.74 (0.66 – 0.83)		
		(LC) Cough	0.58 (0.53 – 0.65)		
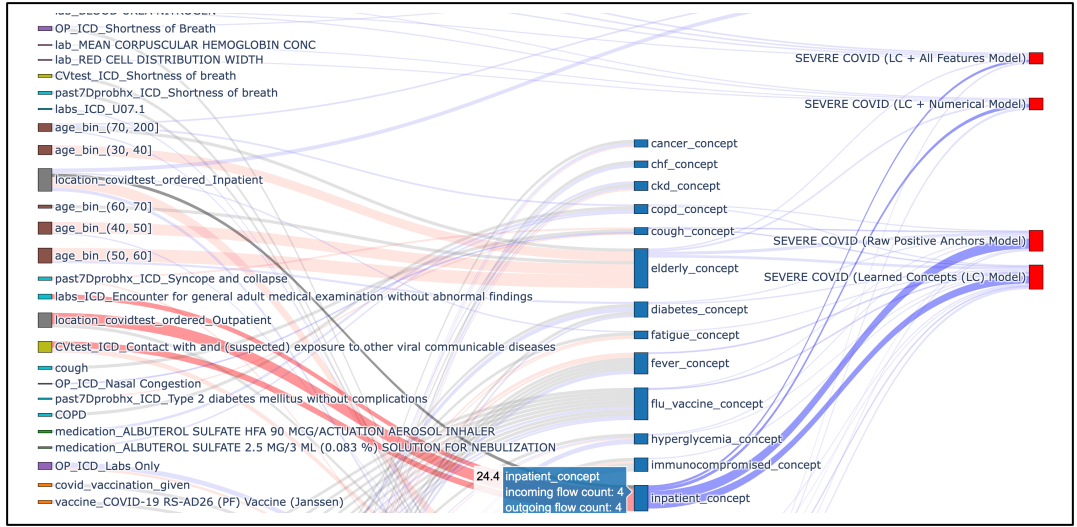
Figure 1. Screenshot of interactive Sankey diagram showing how raw features (first column) translate into clinical concepts (second column), and how both are ultimately used in each model (third column). Magnitude of coefficients correspond to flow thickness, positive log HRs are blue, and negative log HRs are red. Black flows indicate positive anchors for the corresponding concept. Visit acmilab.org/severe_covid to interact with the full diagram.
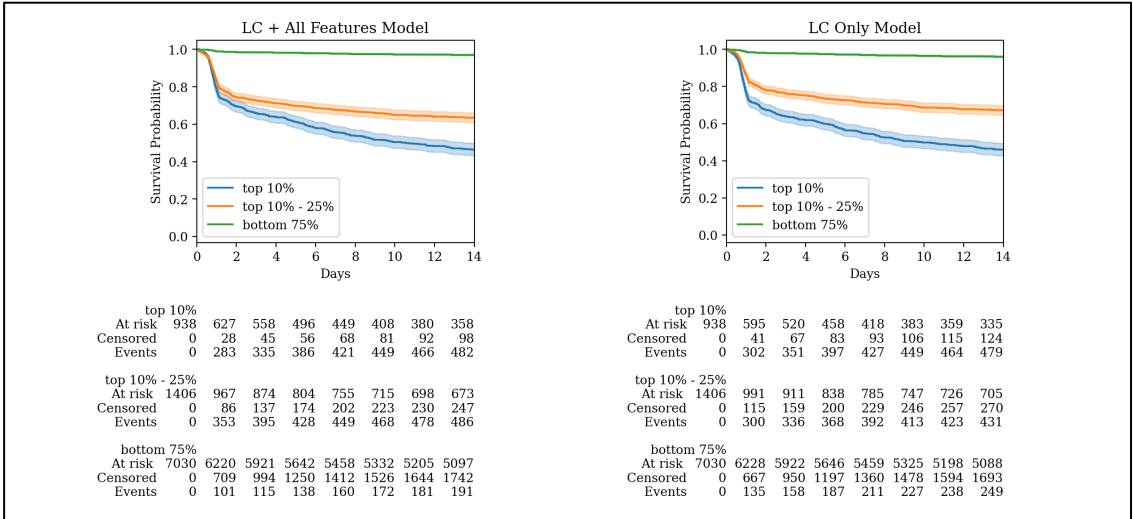

Figure 2. Kaplan Meier survival curves for the high (top 10%), medium (top 10-25%), and low (bottom 75%) risk groups using predictions from the LC + All Features model (left) and the LC only model (right). Counts at the bottom show the number of individuals who are at risk, are censored, or experienced the severe COVID-19 event across time.
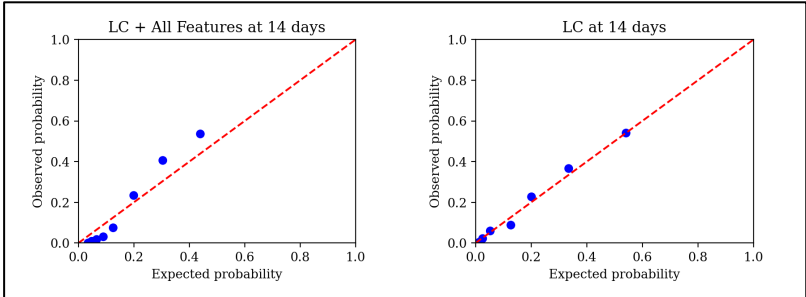

Figure 3. One-calibration of the LC + All Features model (left) and LC model (right) at 14 days, binned into ten groups. Red dotted line corresponds to perfect calibration.
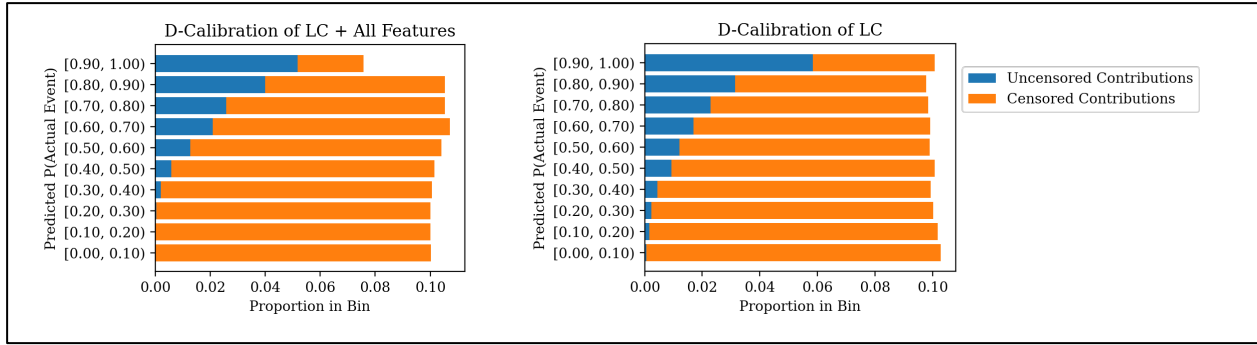
Figure 4. D-calibration histogram of LC + All Features model (left) and LC model (right), binned into ten groups. In a completely D-calibrated model,[18] all of the horizontal bars should be at 0.10.

Table 5. Back-testing performance of All Features, LC + All Features, and LC only over 3-month seasons. Spring (SP) is March 20th until June 21st, followed by summer (SU) until September 22nd, followed by fall (F) until December 21st, followed by winter (W) until March 20th.

All Features only, trained up to:	Test C-index evaluated on:							
	SP 2020	SU 2020	F 2020	W 2020	SP 2021	SU 2021	F 2021	W 2021
End of spring 2020	0.842	0.903	0.855	0.839	0.804	0.841	0.808	0.845
End of summer 2020	-	0.713	0.694	0.697	0.622	0.699	0.667	0.711
End of fall 2020	-	-	0.882	0.868	0.813	0.855	0.84	0.907
End of winter 2020	-	-	-	0.749	0.646	0.718	0.718	0.735
End of spring 2021	-	-	-	-	0.818	0.856	0.844	0.908
End of summer 2021	-	-	-	-	-	0.859	0.847	0.91
End of fall 2021	-	-	-	-	-	-	0.850	0.911
1/12/2022 (study end)	-	-	-	-	-	-	-	0.912

All Features + LCs, trained up to:	Test C-index evaluated on:							
	SP 2020	SU 2020	F 2020	W 2020	SP 2021	SU 2021	F 2021	W 2021
End of spring 2020	0.791	0.840	0.852	0.837	0.774	0.814	0.805	0.876
End of summer 2020	-	0.847	0.852	0.845	0.775	0.806	0.818	0.891
End of fall 2020	-	-	0.852	0.845	0.781	0.812	0.822	0.896
End of winter 2020	-	-	-	0.846	0.778	0.811	0.823	0.896
End of spring 2021	-	-	-	-	0.778	0.813	0.827	0.899
End of summer 2021	-	-	-	-	-	0.812	0.826	0.899
End of fall 2021	-	-	-	-	-	-	0.827	0.900
1/12/2022 (study end)	-	-	-	-	-	-	-	0.900

LCs only, trained up to:	Test C-index evaluated on:							
	SP 2020	SU 2020	F 2020	W 2020	SP 2021	SU 2021	F 2021	W 2021
End of spring 2020	0.797	0.863	0.869	0.850	0.795	0.833	0.83	0.904
End of summer 2020	-	0.857	0.863	0.842	0.785	0.819	0.822	0.904
End of fall 2020	-	-	0.867	0.850	0.800	0.828	0.831	0.909
End of winter 2020	-	-	-	0.854	0.801	0.826	0.834	0.911
End of spring 2021	-	-	-	-	0.803	0.828	0.834	0.911
End of summer 2021	-	-	-	-	-	0.829	0.833	0.910
End of fall 2021	-	-	-	-	-	-	0.837	0.914
1/12/2022 (study end)	-	-	-	-	-	-	-	0.913

Discussion

Learned Concept Classifiers. The strongest coefficients for each concept classifier are often not features that immediately come to mind, but nevertheless match clinical intuition. For the old age concept, we observe that a high PF flu vaccine has a large coefficient (3.31), and is a vaccine only given to patients 65 and older. The outpatient shortness of breath symptom (3.23) is the second highest coefficient for the inpatient concept, possibly indicating that outpatients with this symptom are at high risk of becoming an inpatient. The shortness of breath concept depends on dexamethasone (0.75) which relieves inflammation, and albuterol sulfate (0.67), prescribed for lung conditions. For the obesity concept, sleep apnea, which is often caused by excess weight, has the largest coefficient (0.96).

Note, however, that the learned concepts may not be a perfect representation of the underlying concept. None of the concept classifiers perfectly recover the originally known positives, with recall ranging from 0.381 to 0.974 (Table 2). This could be due to insufficient signal in the remaining covariates or underfitting due to the simplicity of the logistic regression model class. Additionally, some concepts learn substantially more positives than were available in the original data. For example, obesity originally has 433 positives in the data but the concept classifier marks 2,157 patients as having obesity with probability greater than 0.5. It is difficult to verify the faithfulness of the concept classifiers to the true concepts without manual review, but it is possible that the learned concepts may mark patients as "obesity-like" based on their other covariates rather than learning whether they truly have obesity. Additionally, while learned concepts are amenable to interpretation through the coefficients of the concept classifier, they still require domain expertise to manually define positive anchor variables. Finally, the conditional independence assumptions of the selected anchors may not hold in practice, and these assumptions are difficult to verify.

Learned Survival Models. The coefficients for the survival models trained on LCs, All Features, and LC + All Features are mostly consistent with clinical intuition. Inpatients are more likely to experience adverse outcomes than those not hospitalized, old age is well-documented to be associated with higher COVID-19 death rates,[1,7,8] shortness of breath indicates respiratory involvement, and medications such as dexamethasone, acetaminophen, and intravenous saline are given to hospitalized patients. Higher BUN indicates worse liver and kidney function, and COVID-19 vaccines are designed to protect against severe COVID-19. While it is surprising that some of the learned concepts for fever and cough symptoms have negative HRs, upon inspection we find that these are most reliably recorded for outpatients and may encode some additional information about outpatient status. From exploring the interactive visualization, some of the features selected by the All Features model are used by LCs in the LC + All Features model, possibly indicating that they serve as proxies for higher level concepts. For example, the saline IV bolus, present only in the All Features model, is used in the inpatient concept classifier with a positive coefficient. We also note that the coefficients are likely shrunken towards zero due to the Lasso penalty, and the non-informative or independent censoring assumption of the Cox proportional hazards model may not hold since censoring occurs upon discharge.

Our Lasso-Cox models all outperform the baselines (Covichem, Galloway count, Galloway reweighted) in terms of aggregate, inpatient, and outpatient concordance (Table 3). As measured by aggregate concordance, we observe that the learned concepts provide a boost in performance over the raw positive anchors (C-index 0.858 vs. 0.844). This boost in performance places the LC model approximately halfway between the performance of the raw positive anchors and the LC + All Features and All Features models, which both achieve a C-index of 0.872. For LC + All Features and All Features models, the C-index on the outpatient subpopulation (0.879 and 0.880) is higher than that on the entire cohort, whereas the C-index on the inpatient subpopulation (0.715 and 0.717) is lower. For all remaining models, the performance on the inpatient and outpatient subpopulations is lower than in aggregate, possibly indicating that it is easy to order the relative risks of inpatients versus outpatients. The LC model appears slightly better calibrated than the LC + All Features model, but when used to stratify patients into high, medium, and low-risk strata, both models yield groups with clear separation between their survival curves. Finally, while there is some loss of discriminative performance going from the All Features models to the LC only model, when tested under the back-testing framework this gap seems to close quickly on subsequent time periods and the LC model even eventually surpasses the performance of the All Features model. Thus, models with LCs might be more resilient over time than models learned only from All Features. If the set of important high-level concepts themselves change over time, however, new concepts may need to be learned accordingly.

Future Work. In the future, we plan to integrate our models with the healthcare system, and to continue to monitor the performance of the models over time. It will be important to further study how high-level clinical concepts perform across different settings, and to quantify the extent to which classifiers learned on top of these concepts might be transferable across hospitals. As COVID-19 continues to evolve over time, we will also investigate whether new concepts become relevant for prediction.

References

1. Docherty AB, Harrison EM, Green CA, Hardwick HE, Pius R, Norman L, et al. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. BMJ. 2020 May 22;369:m1985.

2. Henry BM, de Oliveira MHS, Benoit S, Plebani M, Lippi G. Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): a meta-analysis. Clin Chem Lab Med. 2020 Jun 25;58(7):1021–8.

3. King JT Jr, Yoon JS, Rentsch CT, Tate JP, Park LS, Kidwai-Khan F, et al. Development and validation of a 30-day mortality index based on pre-existing medical administrative data from 13,323 COVID-19 patients: The Veterans Health Administration COVID-19 (VACO) Index. PLoS One. 2020 Nov 11;15(11):e0241825.

4. Xu W, Sun NN, Gao HN, Chen ZY, Yang Y, Ju B, et al. Risk factors analysis of COVID-19 patients with ARDS and prediction based on machine learning. Sci Rep. 2021 Feb 3;11(1):2933.

5. Zhou H, Cheng C, Lipton ZC, Chen GH, Weiss JC. Mortality Risk Score for Critically Ill Patients with Viral or Unspecified Pneumonia: Assisting Clinicians with COVID-19 ECMO Planning. In: 18th International Conference on Artificial Intelligence in Medicine (AIME) 2020, Proceedings. Springer-Verlag; 2020. p. 336–47.

6. Bats ML, Rucheton B, Fleur T, Orieux A, Chemin C, Rubin S, et al. Covichem: A biochemical severity risk score of COVID-19 upon hospital admission. PLoS One. 2021 May 6;16(5):e0250956.

7. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, et al. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. JAMA Intern Med. 2020 Aug 1;180(8):1081–9.

8. Galloway JB, Norton S, Barker RD, Brookes A, Carey I, Clarke BD, et al. A clinical risk score to identify patients with COVID-19 at high risk of critical care admission or death: An observational cohort study. J Infect. 2020 Aug;81(2):282–8.

9. Salaffi F, Carotti M, Tardella M, Borgheresi A, Agostini A, Minorati D, et al. The role of a chest computed tomography severity score in coronavirus disease 2019 pneumonia. Medicine . 2020 Oct 16;99(42):e22433.

10. Li K, Wu J, Wu F, Guo D, Chen L, Fang Z, et al. The Clinical and Chest CT Features Associated With Severe and Critical COVID-19 Pneumonia. Invest Radiol. 2020 Jun;55(6):327–31.

11. Choi E, Xiao C, Stewart W, Sun J. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. Adv Neural Inf Process Syst. 2018;31.

12. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine. 2021 May 20;4(1):86.

13. Elkan C, Noto K. Learning Classifiers from Only Positive and Unlabeled Data. 2008;

14. Halpern Y, Horng S, Choi Y, Sontag D. Electronic medical record phenotyping using the anchor and learn framework. J Am Med Inform Assoc. 2016 Jul 1;23(4):731–40.

15. Bekker J, Davis J. Learning from positive and unlabeled data: a survey. Mach Learn. 2020 Apr 1;109(4):719–60.

16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(85):2825–30.

17. Davidson-Pilon C. lifelines: survival analysis in Python. J Open Source Softw. 2019 Aug 4;4(40):1317.

18. Haider H, Hoehn B, Davis S, Greiner R. Effective Ways to Build and Evaluate Individual Survival Distributions. J Mach Learn Res. 2020;21(85):1–63.