

# Preparing For The Next Pandemic: Transfer Learning From Existing Diseases Via Hierarchical Multi-Modal BERT Models to Predict COVID-19 Outcomes

**Khushbu Agarwal**

Pacific Northwest National Laboratory

**Sutanay Choudhury** (✉ [Sutanay.Choudhury@pnnl.gov](mailto:Sutanay.Choudhury@pnnl.gov))

Pacific Northwest National Laboratory

**Sindhu Tipimani**

Virginia Tech

**Pritam Mukherjee**

Stanford University

**Colby Ham**

Pacific Northwest National Laboratory

**Suzanne Tamang**

Stanford University

**Matthew Baker**

Stanford University

**Siyi Tang**

Stanford University

**Veysel Kocaman**

John Snow Labs

**Olivier Gevaert**

Stanford University

**Robert Rallo**

Pacific Northwest National Laboratory

**Chandan Reddy**

Virginia Tech

---

**Research Article**

**Keywords:** infectious diseases, COVID-19, biomarkers, risk factors

**Posted Date:** November 12th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-965815/v1>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

# Preparing for the Next Pandemic: Transfer Learning from Existing Diseases via Hierarchical Multi-Modal BERT Models to Predict COVID-19 Outcomes

Khushbu Agarwal<sup>1</sup>, Sutanay Choudhury\*<sup>1</sup>, Sindhu Tipirneni<sup>2</sup>, Pritam Mukherjee<sup>3</sup>, Colby Ham<sup>1</sup>, Suzanne Tamang<sup>4</sup>, Matthew Baker<sup>6</sup>, Siyi Tang<sup>5</sup>, Veysel Kocaman<sup>7</sup>, Olivier Gevaert<sup>3,4</sup>, Robert Rallo<sup>1</sup>, and Chandan K Reddy<sup>2</sup>

<sup>1</sup>Pacific Northwest National Laboratory, Richland, 99354, USA

<sup>2</sup>Department of Computer Science, Virginia Tech, Arlington, 22203, USA

<sup>3</sup>Stanford Center for Biomedical Informatics Research, Department of Medicine, School of Medicine, Stanford University, Stanford, 94305, USA

<sup>4</sup>Department of Biomedical Data Science, Stanford University, Stanford, 94305, USA

<sup>5</sup>Department of Electrical Engineering, Stanford University, Stanford, 94305, USA

<sup>6</sup>Division of Immunology and Rheumatology, Department of Medicine, Stanford University, Stanford, 94305, USA

<sup>7</sup>John Snow Labs, Delaware City, 19958, USA

\*Sutanay.Choudhury@pnnl.gov

## ABSTRACT

Developing prediction models for emerging infectious diseases from relatively small numbers of cases is a critical need for improving pandemic preparedness. Using COVID-19 as an exemplar, we propose a transfer learning methodology for developing predictive models from multi-modal electronic healthcare records by leveraging information from more prevalent diseases with shared clinical characteristics. Our novel hierarchical, multi-modal model (TRANSMED) integrates baseline risk factors from the natural language processing of clinical notes at admission, time-series measurements of biomarkers obtained from laboratory tests, and discrete diagnostic, procedure and drug codes. We demonstrate the alignment of TRANSMED's predictions with well-established clinical knowledge about COVID-19 through univariate and multivariate risk factor driven sub-cohort analysis. TRANSMED's superior performance over state-of-the-art methods shows that leveraging patient data across modalities and transferring prior knowledge from similar disorders is critical for accurate prediction of patient outcomes, and this approach may serve as an important tool in the early response to future pandemics.

## Introduction

The COVID-19 pandemic revealed salient challenges in developing systems that can accurately predict outcomes associated with an emerging infectious disease. In particular, it emphasized the need for hospitals to access risk stratification tools that could be used to proactively identify COVID-19 patients at a greater risk of undesirable outcomes<sup>1-8</sup>. Such capabilities are critical for institutions to prioritize resources, and bring a quantitative approach to triaging<sup>9</sup> in an emergency, which subjects the human caregivers to intense psychological stress. Undertaking hard and pragmatic decisions, and accepting their consequences leads to a new crisis that is appropriately called the "hidden pandemic for healthcare workers"<sup>10-12</sup>. Due to the lack of historical COVID-19 cases for training supervised machine learning models, early methods for COVID-19 severity prediction focused on the analysis of a relatively small number of carefully chosen model covariates, which included demographic risk factors, prior comorbidities, symptoms on admission, and laboratory biomarkers<sup>8,13</sup>. These carefully chosen covariates were predominately used to train multivariate logistic regression and boosted decision tree-based approaches<sup>1,14-16</sup>. Electronic health records are heterogeneous data sources that include unstructured clinical notes, structured data that are coded as ICD diagnoses and CPT procedures, and numeric measurements such as body vitals along with various laboratory test results. Due to their size, richness, and wide-scale adoption, the past few years have seen major progress in developing predictive models for different subsets of such data sources, where deep learning methods have been shown to achieve state-of-the-art results for several medical outcomes such as re-admissions, mortality prediction, and length of stay<sup>13,17-26</sup>. However, a notable gap lies in integrating all of the multi-modal information into a single predictive model, and the challenges are amplified by the need for large amounts of training data. In this paper, we propose TRANSMED, a methodology for developing multi-modal predictive models, while addressing training data scarcity issues posed by emerging (or rare) diseases through transfer learning from diseases with shared

cohort-level characteristics and similar outcomes.

To address the existing gaps in pandemic preparedness, we sought to improve on current methods to: (i) predict if a patient will be staying in the hospital, after a certain time using the patient’s multi-modal history. This provides a better understanding of the severity of a patient’s condition, and (ii) predict the likelihood of a patient requiring mechanical ventilation. Collectively, these prediction tasks capture the inherent challenges of inpatient resource planning such as those to predict which patients are most likely to experience poor outcomes over a span of next 3-7 days<sup>9,27-30</sup>.

Figure 1 presents a case study of a real COVID-19 patient to illustrate how different modalities offer unique information to reason about a patient’s current state and future evolution. However, making use of the information requires careful handling of sparsity across time and data sources. Diagnostic codes provide a more definitive assertion of patient’s short- and long-term medical conditions, but they do not provide continuous observation of the patient. In addition, they are not frequently observed and may miss key signals related to clinical deterioration. Observing the occurrence of key procedures and laboratory measurements provide clinical information on a patient’s immediate conditions. The drug data stream provides additional treatment information that helps characterize a patient’s disease state; compared to clinical observations such as diagnostic codes, procedures, and laboratory tests, which typically marks the onset of a problem, the duration of a medication allows us to reason about the type and severity of a particular symptom and its short- and long-term consequences; thereby serving as a bridge to connect other data sources that more sporadically report information.

**Our contributions.** TRANSMED uses a hierarchical approach for learning cross-modal interactions across medical concepts that occur closely in time. The self-supervised methodology implemented via BERT layers<sup>31</sup> is first used to learn the higher-order fine grained medical concept interactions using a Severe Respiratory Disease (SRD) cohort from Stanford Hospital with 9,348 patient hospitalizations. Next, specific layers of the model are further trained for modeling temporal trajectories of COVID-19 patients using EHR data of 1,701 patients. Our neural architecture is distinct from the other recent BERT-based<sup>31,32</sup> prediction models in multiple dimensions. Our model integrates temporal information in multiple representations, that includes clinical notes, discrete entity-based representation of diagnostic codes, drug codes, procedure codes along with continuous valued time-series measurement of laboratory tests. To the best of our knowledge, the proposed model demonstrates the widest integration of EHR-based data sources across multiple modalities for COVID-19 severity prediction. Our experiments show that our hierarchical transfer learning based approach using Severe Respiratory Disease (SRD) cohorts leads to an average improvement of 12.9% and 10.3% in AUROC for COVID-19 patient stay and ventilation prediction. We benchmark our implementation with three models representing distinct prediction approaches and demonstrate an improvement ranging over 5.8%-29.2% for AUROC and 3.6%-66% in F1 score measure for ventilation prediction tasks, and accurately predicting the likelihood of short- and long-term patient stays.

We also present a new methodology to interpret and evaluate model predictions via multi-comorbidity analysis. Much of the literature on EHR models focus on characterizing the prediction performance through univariate analysis of the well understood risk factors<sup>29,33</sup>. However, clinical presentation of a disease and its severity can markedly vary depending on the constellation of symptoms, prior health conditions and risk factors. Deep learning models are adept at learning higher-order feature interactions. Therefore, evaluating model recommendations solely in terms of single factors may not explain when a prediction is driven by a hidden combination of multiple factors. We present a methodology for identifying top multi-comorbidity conditions in a data-driven fashion and evaluate their relative impact on model predictions. We believe our analysis would motivate data-driven discovery of key multi-comorbidities associated with a disease while advancing the interpretability and rigor for evaluating deep learning models for clinical use.

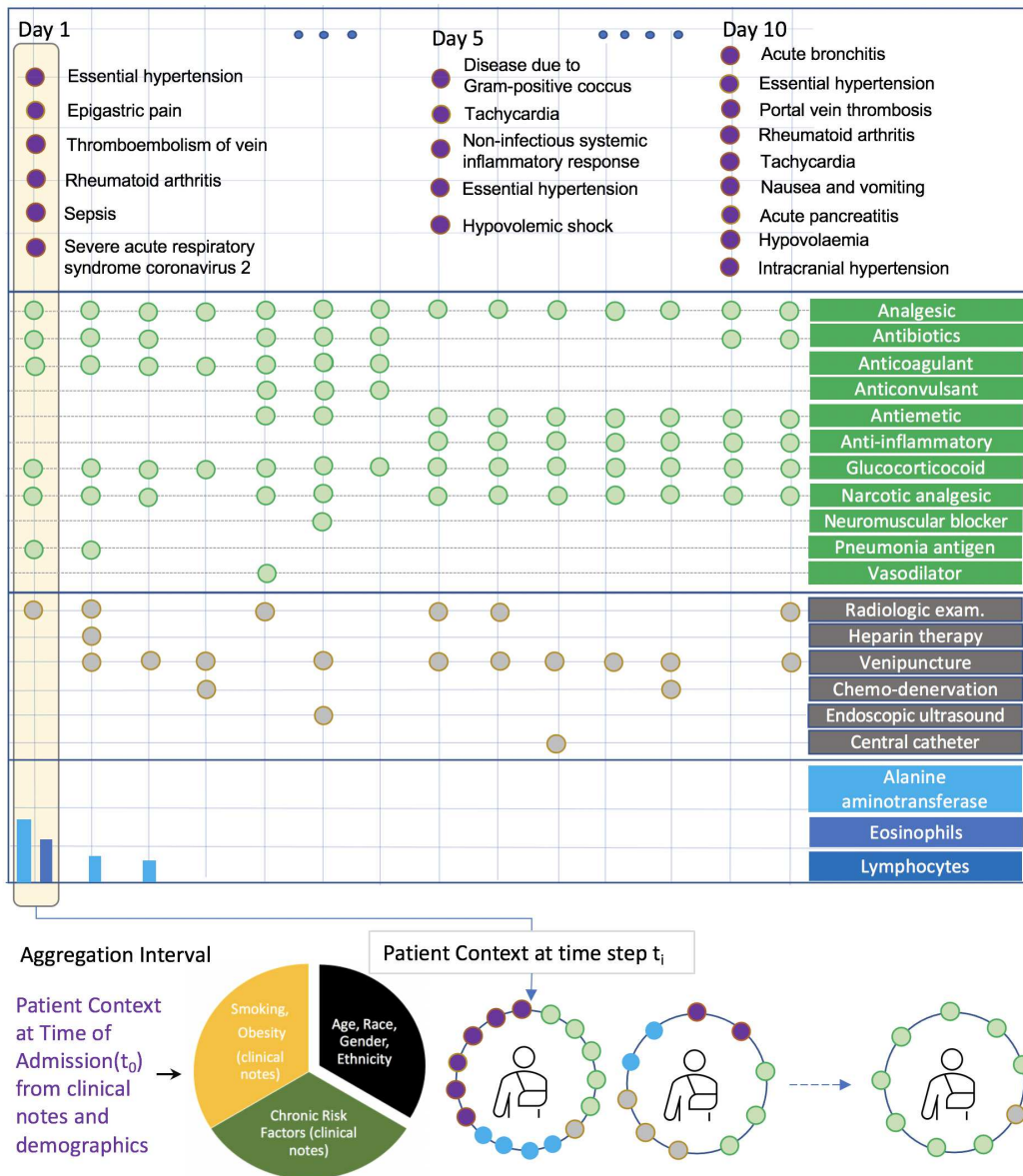
## Methods

We begin this section with a description of the available data sources and the cohort selection process. After that, the problem statement is described followed by a description of the model architecture.

### Data Sources

Our study is based on de-identified EHR data of all patients treated at Stanford Hospital, between January 1, 2015 and March 19, 2021. This dataset was provided via STanford Research Repository (STARR)<sup>34</sup> and was used under approval by Stanford University Institutional Review Board (IRB) protocol: 50033 (Machine Learning of Electronic Medical Records for Precision Medicine). Patient informed consent was waived by Stanford University Institutional Review Board (IRB) for this protocol. All methods were carried out in accordance with relevant guidelines and regulations.

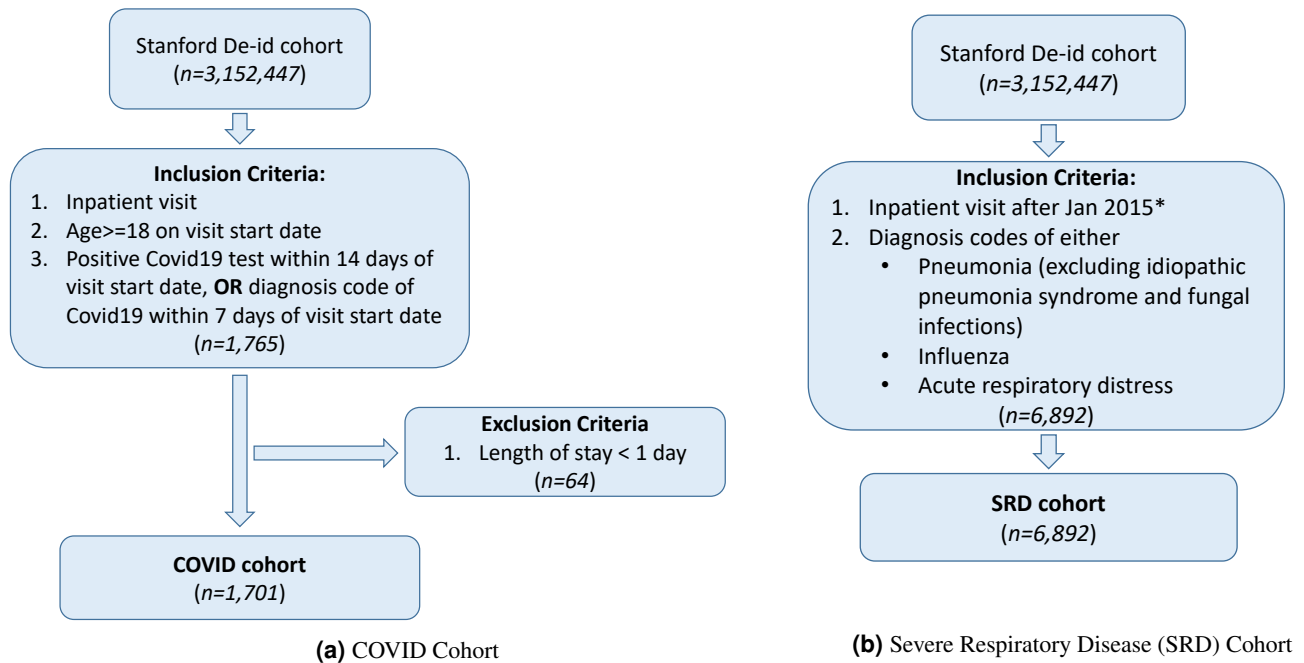
As part of the de-identification process, the actual admission dates were randomized up to 30 days. The data was retrospectively collected during the practice of care and transformed into the OMOP Common Data Model Version 5, (<https://www.ohdsi.org/data-standardization/the-common-data-model/>) by the STARR OMOP team. Using the STARR OMOP data, we created our cohort of COVID-19 patients (Figure 2) based on the following inclusion criteria: 1) patients with inpatient visits after January 1, 2020, 2) patient age greater than 18 at admission, 3) patient had either a positive COVID-19 test within



**Figure 1.** An illustration of multi-modal data sources observed over the course of a COVID-19 patient's stay in the hospital. The colors indicate diagnosis (purple), drugs (green), procedures (gray), and numeric lab measurements (blue bars). Different data modalities are observed at varying frequency in raw patient data, with lab measurements being the most sparse across patients and across time. TRANSMED reduces the impact of sparsity by utilizing all modalities of data in a given time interval (e.g., 24 hours), creating more informed patient state snapshots in time.

14 days prior to the admission or had a diagnosis of COVID-19 within 7 days prior to the admission. Visits that were less than 1 day in length were excluded. We also created a cohort that included hospital admissions for severe respiratory disease (SRD) patients with influenza, pneumonia or ARDS, for our transfer learning approach. The ICD-9 codes were first mapped to OMOP CDM V5 concept identifiers which were then used to execute queries to retrieve the cohort data. The specific codes used for the cohort selection and mechanical ventilation are listed in the Supplementary Table 1. The cohorts were cross-referenced for similarity in observed medical codes and ventilation outcomes. Table 1 provides the summary statistics for the two cohorts, with a detailed comparison under the results section.

Our study cohorts include clinical observations from four data sources: 1) free-text patient notes at the time of admission, 2) discrete codes (we also refer to them as structured data) representing diagnosis codes, prescribed drugs, laboratory tests that were ordered, and codes for the procedures performed, 3) continuous time-series measurements that were available for a subset



**Figure 2.** Cohort selection process for COVID-19 and Severe Respiratory Disease (SRD) patients from the Stanford Hospital. \*For the SRD cohort, the start year 2015 was chosen heuristically to ensure that the data was sufficient in terms of both quantity and quality.

of the ordered laboratory tests, and 4) patient demographics (age/race/sex/ethnicity). We parsed each note into sections and used the SparkNLP library<sup>35</sup> named entity recognizer (NER) for extracting medical conditions from the clinical notes (see Supplementary section on “Data Sources” for implementation details). The extractions were used to determine the presence or absence of baseline risk factors for each patient at the time of admission, including: Coronary Artery Disease (CAD), diabetes, family history, hyperlipidemia, hypertension, existing medication, obesity, and smoking. Note that most of the COVID-19 patients were de novo, without any prior medical records. Family history and existing medication were discarded due to insufficient coverage.

### Problem Statement

Let  $X_t = (C_t, M_t)$  be the patient state at time  $t$ , where  $C_t \subseteq \mathcal{C}$  is a set of observed codes,  $M_t \in \mathbb{R}^{|M|}$  is a vector of lab values, and  $\mathcal{C}$  and  $|M|$  denote set of codes and lab values observed across the whole cohort respectively. Let  $d \in \{0, 1\}^{|D|}$  and  $r \in \{0, 1\}^{|R|}$  be multi-hot vectors denoting patient demographics and risk factors obtained at the time of admission, respectively. The clinical outcome at time  $t$  is denoted by  $O_t$ . The problem statement is as follows: Given demographics  $d$ , risk factors  $r$ , and a sequence of  $T_h$  historical states  $X_{t-T_h+1}, X_{t-T_h+2}, \dots, X_t$ , predict a clinical outcome of interest  $T_f$  steps ahead into future, denoted by  $O_{t+T_f}$ . Following the tasks proposed in the introduction section, we will focus on two outcomes: patient staying at the hospital or patient requiring mechanical ventilation at time  $(t + T_f)$ .  $T_h$  and  $T_f$  are referred to as “# input time-steps” and “look-ahead”, respectively.

### Model Architecture

In this section, we present the intuition behind the key components of our model and provide details of the proposed hierarchical model architecture. The supplementary section titled “Related Work” provides a detailed overview of all prominent methods that has been applied over diverse modalities and prediction methodologies for EHR data. Figure 3 presents the data-flow through our hierarchical multi-modal model with two primary components: 1) Transfer learning driven top layers that accept patient state at a given time (subsequently referred to as “patient context”) as input and produces a contextualized representation, capturing feature interactions across all modalities of data in a single time step. 2) The bottom temporal modeling layer to model patient evolution over time. This layer takes as input, the contextualized vector representations output from top layers for each time interval augmented with positional encoding (indicating the time of observation) and produces a vector representing patient evolution over time.

The transfer learning component is implemented using BERT layers due to their proven effectiveness in learning contextual

relationships between set of observations<sup>31</sup>. The choice for specific implementation of the temporal modeling layer is flexible, and we discuss the different options for this layer in detail below. Overall, we use the SRD cohort to train the cross-modal interaction layers during transfer learning. We train the temporal layers on the COVID-19 patient data only and keep the multi-modal transformer layers fixed from the training of SRD cohort.

**Multi-Modal Patient Context Encoders** We begin with describing the “patient context” data structure that captures the state of a patient over a single aggregation interval, across all modalities of patient data. We define the patient context as a collection of labeled sets that contains both static attributes (such as demographics and risk factors) and all multi-modal EHR information available during a specific interval in time.

We map numeric age values into 11 bins, based on CDC criteria (<https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html>). Age, race, and ethnicity are represented as categorical variables, while the rest of the variables such as sex and presence of risk factors are represented as boolean variables (shown in Figure 3a). The temporal information associated with the patient context at a given time interval can be naturally recognized as a collection of sets, where each set represents a type of information such as diagnosis codes, drugs, procedures, laboratory tests and contains the discrete identifiers of associated diagnosis codes, drugs, etc. (shown in Figure 3b). The multi-modal patient context encoder module takes in the collection of sets as described above and returns a single vector representing the patient context. We encode each boolean and categorical variables as a one-hot vector. The last component in the patient context are laboratory measurements, which are represented as a dictionary of key-value pairs representing a test and it’s associated numeric value. We only used the laboratory measurements from the COVID-19 cohort due to unit consistency issues in the SRD data. A concatenation of all these vectors yields the final embedding, representing the patient context.

**Learning Contextual Representation of Patient State in a Time Interval.** Our first component focuses on learning a contextualized representation of the patient state per time interval. This *contextualization* is important. For example, a drug may be recommended as an “if-then” measure, where it is used if the patient descends into a critical condition. If a physician were to read a chart with such information, they would understand this series of events and recognize why the drug exists on the patient record. On the contrary, if the patient was already in a condition where the drug had to be administered, it would be reflected through the presence of other diagnostic codes and lab measurements. The BERT layer in Figure 3 accomplishes this contextualization effect, and generates a different vector embedding for each entity (such as the drug) depending on the input patient context.

**Transfer Learning** We adopt a self-supervised learning approach<sup>36,37</sup> to train the BERT-based multi-modal patient context encoder layer (green box in Figure 3c). Given a patient context  $C_t = (c_t^1, \dots, c_t^{|C_t|})$ , we generate a random mask  $m_t \in \{0, 1\}^{|C_t|}$  to replace a specified number of condition, drug or procedure codes with a special [MASK] token. and train the model to predict the missing values from the rest of the patient context using categorical cross-entropy loss.

$$\hat{C}_t = (\hat{c}_t^1, \dots, \hat{c}_t^{|C_t|}) \quad \text{where} \quad \hat{c}_t^i = \begin{cases} [MASK], & \text{if } m_t[i] = 1 \\ c_t^i, & \text{otherwise} \end{cases} \quad (1)$$

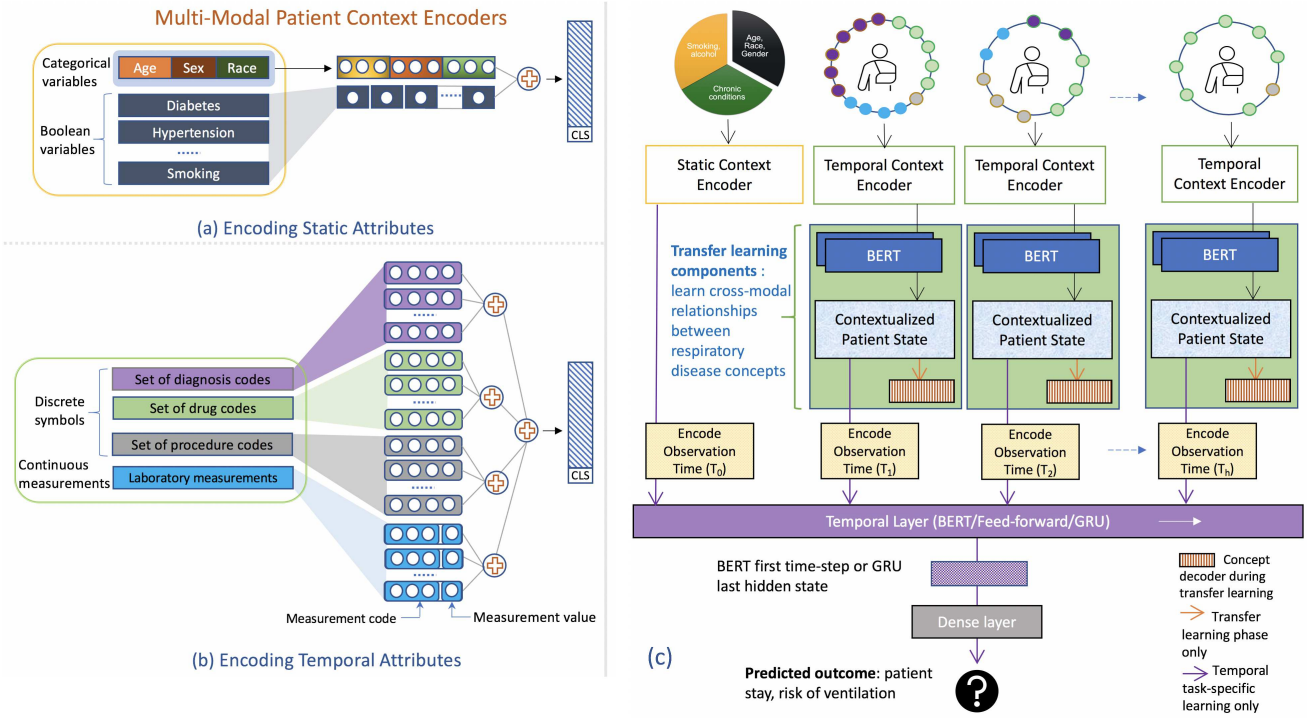
Unlike existing self-supervised methods for EHR data<sup>19,31,32</sup> that mask and predict codes encoded within the complete patient trajectory, we perform contextualization within a small time step (e.g., 12 to 24 hours) without encoding time. This facilitates learning a more fine-grained contextualization model from SRD, which is essential for transfer learning across diseases where the patient conditions may evolve at different time-scales<sup>7,14</sup>. The masked patient context  $\hat{C}_t$  is then passed through a BERT module to get contextualized patient state at time  $t$ .  $h_t^C$  denotes the embedding for the entire context.

$$h_t^C, h_t^1, \dots, h_t^{|C_t|} = BERT(\hat{c}_t^1, \dots, \hat{c}_t^{|C_t|}) \quad (2)$$

which are passed through a linear layer to predict the masked codes.

$$\tilde{c}_t^i = \text{softmax}(w^T h_t^i) \in \mathbb{R}^{|\mathcal{C}|} \quad \forall i = 1, \dots, |C_t| \quad (3)$$

**Learning over time from the entire history.** Finally, the temporal modeling step introduces additional layers on top of the pre-trained layers to specialize the model for specific prediction tasks. All encoded patient contexts ( $h_i^C, i = 1, \dots, T_h$ ) are further augmented with a relative positional encoding ( $r_i^{pos}$ )<sup>36</sup> to produce the inputs (denoted as  $\hat{X}_i$  below) going into the temporal layer (purple box in Figure 3c). The position encoding represents the offset in time as measured from the beginning of the patient stay, and allows us the model to reason about variable gaps in the patient data. For example, assuming we aggregate a patient’s information using a 24 hour interval, and we have 3 diagnosis codes being reported at 3 PM (day 1), 9 AM (day 3), and 3 AM (day 4), they would be associated with a positional index of 1, 3, and 4. For the static patient attributes, we encode them as shown in Figure 3a and pad appropriately to produce a vector of same size as the  $\{h_i^C\}$ , and further combine with the positional offset of zero to produce  $\hat{X}_0$ .



**Figure 3.** TRANSMED architecture. Patient context encoders are shown in (a) for static attributes and (b) for multi-modal temporal attributes. The proposed hierarchical transfer learning model is shown in (c). The transfer learning components take as input the patient’s multi-modal encoded state and produce a contextualized vector. The vectors for all time steps are combined along with static attributes to model patient’s (task-specific) evolution over time.

$$H' = f_{select}(GRU(\hat{X}_0, \hat{X}_1, \dots, \hat{X}_{T_h})), f_{select} \text{ returns the last element from GRU output sequence.} \quad (4)$$

$$H' = f_{select}(BERT(\hat{X}_0, \hat{X}_1, \dots, \hat{X}_{T_h})), f_{select} \text{ returns the first element from BERT output corresponding to "CLS".} \quad (5)$$

$$H' = FFN(\hat{X}_0 \oplus \hat{X}_1 \oplus \dots \oplus \hat{X}_{T_h}), \text{ where } \oplus \text{ represents tensor concatenation operation} \quad (6)$$

$$\hat{o}_t = FFN(H') \quad (7)$$

Finally, the resultant sequence from static patient attributes and the encoded vector embeddings from each time step ( $\hat{X}_0, \hat{X}_1, \dots, \hat{X}_{T_h}$ ) are fed into a *temporal layer* (denoted as  $f_T(\cdot)$ ). We experimented with  $f_T(\cdot)$  using the following: a GRU (Gated Recurrent Unit)<sup>38</sup> layer, a feed forward network (FFN), and a multi-ahead attention layer<sup>36</sup>. The temporal layer is followed by a dense layer, and the output from the dense layer is passed through a non-linear function that produces the final scalar output. Handling of the input ( $\{\hat{X}_i\}$ ) and the output from  $f_T(\cdot)$  changes depending on the specific choice of the temporal layer, and the specific implementations are described by equations (4)-(6). We use the binary cross-entropy as the loss function for training the temporal prediction layers combining  $f_T(\cdot)$  and the final dense layer.

## Results

### Comparison of SRD and COVID-19 Cohorts

Our motivation to adopt a transfer learning approach for training a COVID-19 outcome prediction model was inspired by the strong similarity between the vocabularies in SRD and COVID-19 cohorts (Table 1). 88.8% of the diagnosis codes in our COVID-19 cohort were also found in the SRD cohort. Similarly, the drugs, procedures and laboratory measurement codes in the COVID-19 cohort have an overlap of 94.04%, 63.2% and 90.57%, respectively. The SRD cohort also provided a strong coverage for patients with severe outcomes, even though the distributions of outcomes are slightly different from the COVID-19 cohort (Fisher’s exact test,  $P < .05$ ). 14.6% of the patients in the SRD cohort required ventilation as compared to 11.4% in the COVID-19 cohort. Proportion of ICU days and mortality incidents in the SRD cohort are 23.4% and 8.34%, compared to 5.76% and 6.46% in the COVID-19 cohort, respectively.



**Table 1.** Summary of the COVID and SRD cohorts.

	<b>COVID</b>	<b>SRD</b>	<i>P</i>
<b>Number of patients</b>	1701	6892	
<b>Number of hospitalizations</b>	1701	9348	
<b>Quarter-wise distribution of hospitalization</b>			
Before 2020	0	8574	
2020-Q1	35	559	
2020-Q2	329	202	
2020-Q3	350	13	
2020-Q4	566	0	
2021-Q1	421	0	
<b>Length of stay, median (Interquartile range)</b>	4.8 (2.8 – 8.8)	5.0 (2.0 – 13.0)	.42
<b>Age at encounter, mean (SD)</b>	56.8 (18.6)	38.14 (30.9)	<.001
<b>Age at encounter among adults, mean (SD)</b>	56.8 (18.7)	60.11 (18.4)	<.001
<b>Age groups</b>			<.001
<18	0 (0%)	3689 (39%)	
18 –30	166 (10%)	492 (5%)	
30 – 65	903 (53%)	2534 (27%)	
>=65	632 (37%)	2633 (28%)	
<b>Race</b>			<.001
White	582 (34%)	3522 (51%)	
Black or African American	70 (4%)	314 (5%)	
Asian	228 (13%)	1157 (17%)	
American Indian or Alaskan Native	10 (1%)	21 (0%)	
Native Hawaiian or Other Pacific Islander	46 (3%)	173 (3%)	
Other/Unknown	765 (45%)	1705 (25%)	
<b>Ethnicity</b>			<.001
Hispanic or Latino	691 (41%)	1687 (24%)	
Not Hispanic or Latino	989 (58%)	5099 (74%)	
Other/Unknown	21 (1%)	106 (2%)	
<b>Sex</b>			<.001
Male	854 (50%)	3108 (45%)	
Female	847 (50%)	3784 (55%)	
<b>Clinical Outcomes</b>			
ventilation (yes/no)	194 / 1507	1365/7983	<.001
ICU admissions (yes/no)	98 / 1603	2188/7160	<.001
mortality (died/survived)	110 / 1591	780/8568	.01
<b>Input Codes, [common codes/COVID total], [SRD total]</b>			
Diagnosis	2310/2599	6293	
Procedure	1204/1905	5778	
Drugs	2147/2283	4592	
Lab Measurements	1355/1496	2431	

In terms of demographics, the COVID-19 cohort is quite different from the SRD cohort. In particular, the age distribution of patients in the two cohorts is significantly different (chi-squared  $P < .001$ , see Table 1). This is partly due to the fact that we restricted the COVID-19 cohort to adults only. However, if we restrict the analysis to adults only in both cohorts, the difference persists – while the proportion of patients in the age group 18 – 30 is quite similar (9.8% and 8.7% in COVID-19 and SRD cohorts, respectively), the COVID-19 cohort has a significantly higher proportion of patients in the 30 – 65 age group (53.1%) than the SRD cohort (44.8%) (chi-squared  $P < .001$ ), and consequently, the mean age among adults in the COVID-19 cohort is significantly less than in the SRD cohort (Mann-Whitney U-statistic  $P < .001$ ). The cohorts are significantly different with respect to sex (Fisher exact test  $P < .001$ ), with the COVID-19 cohort being more balanced than the SRD cohort. There are significant differences between the cohorts with regard to race as well (chi-squared  $P < .001$ , ignoring the "Other/Unknown" class). Finally, the "Hispanic or Latino" ethnicity is significantly over-represented in the COVID-19 cohort compared to the SRD cohort (chi-squared  $P < .001$ ).

## Training and Evaluation Setup

As mentioned earlier, we evaluate the model on two binary classification tasks on the COVID-19 dataset. To generate input dataset per patient, first, each patient’s stay duration is segmented into intervals of fixed length (*aggregation windows*) and the visit data within each interval is aggregated. For interpretability reasons, we use a 24-hour aggregation interval. We then use the sliding window approach to generate individual samples for all the models by considering each timestep in the visit as the current timestep. No timesteps containing or following the first occurrence of a positive outcome can be part of an input.

We benchmark our model against three methodologies: logistic regression (LR)<sup>16</sup>, a Gated Recurrent Unit (GRU)-based approach<sup>17</sup>, and BEHRT<sup>31</sup> which is a state-of-the-art extension of BERT<sup>36</sup> models for electronic healthcare records. Similar to BEHRT, MedBERT<sup>32</sup> proposed training at patient visit sequence level, and uses only the diagnosis codes from a patient cohort. Hence, we only empirically compare with BEHRT which uses a much wider scale of structured EHR data and is more suitable for the in-stay patient study. We did not consider any time-series model due to the sparse and highly irregular nature of the time-series based laboratory measurements. The COVID-19 cohort was split randomly into 60% train, 20% validation, and 20% test dataset for evaluation purposes. The same split of patient cohorts was used for evaluating all of the methods. The models are evaluated using the AUROC (Area under Receiver Operating Characteristic curve) and the F1-score measure. The F1-score captures both the precision (positive predictive value) and recall (aka sensitivity, the fraction of relevant instances correctly retrieved) capability of the model. If TP, FP, and FN indicate “True Positive”, “False Positive” and “False Negative”, respectively, F1-score measure is given by:

$$recall = \frac{TP}{TP + FN}, \quad precision = \frac{TP}{TP + FP}, \quad F1 = \frac{2 \times recall \times precision}{recall + precision}$$

For each outcome task (the likelihood of the patient staying in the hospital and ventilation risk) we study two variations of prediction into the future: predict short-term (3 days) and long-term (7 days) ( $T_f = 3, 7$ ) patient outcomes. For all variations, we feed 2 days of patient’s past history ( $T_h$ ) to the model. A detailed performance analysis of TRANSMED as a function of input history size and look-ahead duration is provided in Supplementary Table 4.

## Implementation Details

**Baselines:** For LR, we use an aggregated view of patient’s diagnoses, procedures, drugs, lab codes, lab measurements, demographics, and risk factors over time. The model is trained using all variables encoded with the one-hot encoding scheme. BEHRT model adapts the BERT layers for structured EHR data. It considers the sets of medical codes occurring across multiple visits of a patient as a single instance of training sample. Analogous to the NLP domain, each code is embedded similar to a word, and a time-offset embedding is added for each token depending on the visit id. To ensure a fair comparison, the BEHRT model is pre-trained on SRD and fine-tuned on COVID-19 dataset as well. For the GRU baseline, we encode each input timestep as a multi-hot vector of diagnosis, procedure, and medication codes. Additional implementation details for the baseline methods can be found in Supplementary section titled “Methods: Implementation details.”

**TRANSMED:** The BERT encoder layers in our model are implemented using the PyTorch BERT implementation available from Huggingface <https://github.com/huggingface/transformers> and used 2 layers and 2 heads with a hidden size of 64 for most configuration (refer to Supplementary section titled “Methods: Implementation details” for hyperparameter search). We noticeably used a low number of parameters in the model to ensure training convergence with limited data. We mask and predict one token for every patient state input to the model. The model was trained for a maximum of 300 epochs (both at visit level and patient temporal level) or if the validation loss stopped reducing for 15 consecutive epochs. Training was performed using a single NVIDIA Tesla V100 GPU of 16 GB memory capacity, leading to average training time of 2 hours for the transfer learning step and 30 minutes for the temporal modeling layer using COVID-19 cohort. For reproducibility purposes, our code will be made publicly available upon the acceptance of this paper.

**Table 2.** Performance comparison of TRANSMED with other methods. The methods are evaluated for predicting patient stay and ventilation risk in short-term (next 3 days) as well as long-term (next 7 days). TRANSMED’s best performance was observed using BERT as fine tuning layer for long term ventilation prediction while feed-forward layer did best for other tasks.

(a) Patient stay prediction results.

Model	3 days		7 days	
	AUROC	F1	AUROC	F1
LR	0.79 (0.77 - 0.81)	0.67 (0.65 - 0.71)	0.74 (0.71 - 0.77)	0.68 (0.65 - 0.72)
BEHRT	0.68 (0.64 - 0.71)	0.43 (0.42 - 0.44)	0.62 (0.58 - 0.65)	0.43 (0.42 - 0.44)
GRU	0.84 (0.81 - 0.86)	<b>0.77 (0.75 - 0.79)</b>	0.80 (0.76 - 0.83)	0.67 (0.61 - 0.72)
TRANSMED	<b>0.84 (0.82 - 0.86)</b>	0.72 (0.70 - 0.74)	<b>0.80 (0.77 - 0.83)</b>	<b>0.73 (0.70 - 0.76)</b>

(b) Ventilation risk prediction results.

Model	3 days		7 days	
	AUROC	F1	AUROC	F1
LR	0.64 (0.52 - 0.78)	0.31 (0.28 - 0.35)	0.68 (0.56 - 0.80)	0.31 (0.28 - 0.34)
BEHRT	0.63 (0.60 - 0.66)	0.43 (0.42 - 0.44)	0.66 (0.63 - 0.69)	0.2 (0.19 - 0.22)
GRU	0.62 (0.48 - 0.77)	0.5 (0.5 - 0.5)	0.72 (0.59 - 0.87)	0.51 (0.49 - 0.54)
TRANSMED	<b>0.83 (0.77 - 0.89)</b>	<b>0.52 (0.49 - 0.56)</b>	<b>0.77 (0.67 - 0.87)</b>	<b>0.53 (0.49 - 0.57)</b>

## Performance Analysis

We perform extensive experiments to allow us to answer following major questions: 1) which method is the best modeling approach for a specific prediction task? 2) What is the impact of using transfer learning for predicting COVID-19 patient outcomes? 3) How effective are different modalities of data in capturing patient state over various complexities of prediction objective?

Table 2 provides a comparison of TRANSMED with respect to other benchmarks on all prediction tasks. We experimented with different combinations of input data sources for TRANSMED, logistic regression, GRU, and BEHRT, and report the best performance for each model. All studies are performed in sliding window setting as described in the evaluation setup. A primary observation from Table 2 is that the performance gaps between the benchmarked methods vary depending on the prediction task. Arguably, the patient stay prediction is a simpler task since it requires developing a coarser-level understanding of patient’s severity. A patient can stay in the hospital for a variety of reasons and learning the association between all potential factors and a severity level is key to accurate prediction. On the other hand, predicting a patient’s ventilation risk requires reasoning about a more specific set of symptoms. Also, accurate prediction of ventilation risk requires robustness against label imbalance due to the rareness of the outcome. TRANSMED performs on par with GRU for the patient stay prediction task and consistently outperforms logistic regression and BEHRT. For the two ventilation prediction tasks, TRANSMED demonstrates an average improvement of 17.5% for AUROC and 34.84% for F1 measure over logistic regression, the next best performing method. A detailed comparison of the AUROC profiles for TRANSMED and other baselines is provided in Supplementary Figure 1.

**Impact of transfer learning and hierarchical model** With regards to self-supervised learning approaches, TRANSMED demonstrates an average gain of 25.6% for AUROC and 45.8% for F1 measures over BEHRT across four tasks. The significant out-performance of TRANSMED over BEHRT demonstrates the impact of our hierarchical modeling approach beyond adapting BERT and pre-training on multiple cohorts. The impact of transfer learning and multi-modality is studied in Table 3. Transfer learning makes a significant impact in improving performance across all four tasks, with an average improvement of 12.9% and 10.3% in AUROC for patient stay and ventilation prediction, respectively. Noticeably, transfer learning is the primary contributing factor in the significant performance difference between TRANSMED and other baseline methods for the ventilation prediction tasks. This demonstrates transfer learning helps learn the fine grained interactions between medical concepts that are essential to accurate prediction of complex medical outcomes, in presence of limited training data. For multi-modality, we see that, using only the procedure codes can offer significant predictive performance for all methods

**Table 3.** Ablation study results of the proposed TRANSMED model analyzing the impact of transfer learning and data modalities on the final performance. Ablation study was performed with a fixed set of hyper-parameters and feed forward network fine tuning layer. See Supplementary “Methods: Implementation details” for description of hyperparameter tuning.

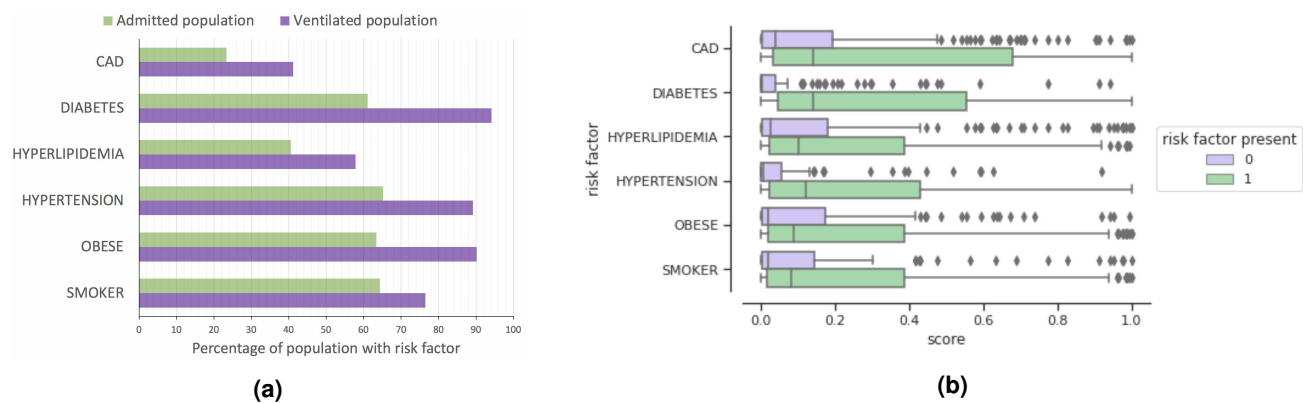
Method	Short-term patient stay		Long-term patient stay		Short-term ventilation		Long-term ventilation	
	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1
<b>Impact of Transfer Learning</b>								
TRANSMED <sub>w/o TL</sub>	0.71	0.62	0.72	0.67	0.72	0.47	0.71	0.48
TRANSMED <sub>with SRD</sub>	0.81	0.69	0.78	0.70	0.79	0.50	0.74	0.49
TRANSMED <sub>with SRD+COVID</sub>	0.83	0.71	0.79	0.72	0.84	0.52	0.77	0.51
<b>Impact of Data Modalities</b>								
TRANSMED <sub>struct</sub>	0.82	0.70	0.76	0.69	0.81	0.54	0.74	0.50
TRANSMED <sub>struct+NLP</sub>	0.83	0.71	0.77	0.69	0.79	0.53	0.75	0.51
TRANSMED <sub>struct+NLP+demo</sub>	0.83	0.72	0.77	0.70	0.81	0.53	0.75	0.51
TRANSMED <sub>struct+NLP+demo+lab</sub>	0.83	0.71	0.79	0.72	0.84	0.52	0.77	0.51

for the patient stay prediction task (see Supplementary Table 5). Introduction of risk factors from clinical notes, laboratory measurements and demographics however improves the accuracy of ventilation prediction by 3.7% and 4.0% in AUROC, respectively. The combination of demographics, laboratory measurements, clinical notes and procedure codes consistently produces the best performance for TRANSMED across four prediction settings. Supplementary Table 5 provides the details for all input variations that were experimented with and reports the performance breakdown for TRANSMED as a function of input features.

### Model Interpretability via Single and Multicomorbidity Analysis

This section presents two different evaluations of our model’s ability to account for important clinical factors. First, we begin with profiling the model’s predicted risk score distribution in terms of well-established univariate risk factors in the clinical literature<sup>28,29</sup>, namely the chronic conditions extracted from clinical notes and demographics information such as age, gender, and sex. Next, we examine the ability of the model to account for important multi-comorbidities. Taking inspiration from the literature on recommendation systems<sup>39</sup>, we demonstrate a new methodology to identify top multi-comorbidities present in the patient population. We then compute a ranked list of the top multi-comorbidities associated with patients with a ventilation outcome, and compare the top-ranked multi-comorbidities as predicted by the model. This is a significantly harder test and we demonstrate that TRANSMED successfully ranks the majority of the multi-comorbidities using this top-k verification approach.

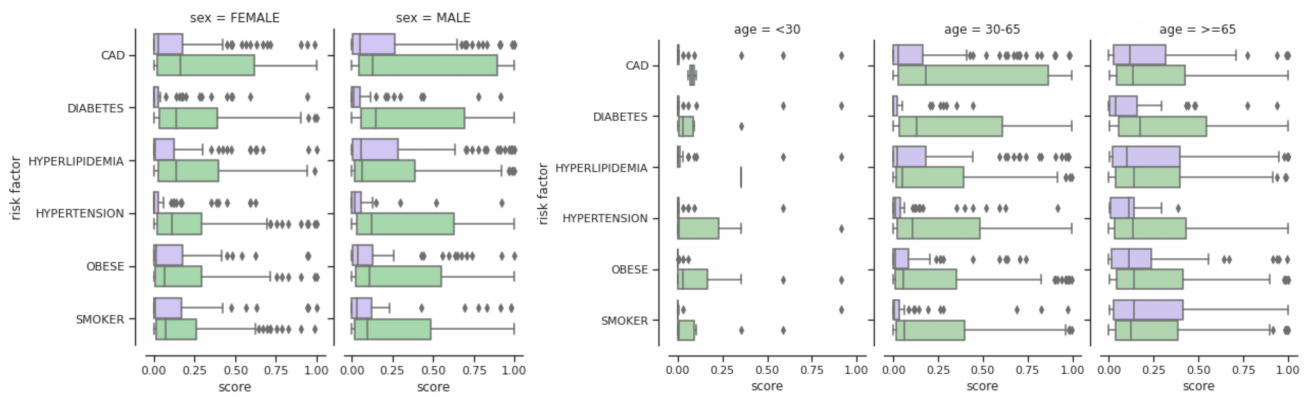
**Defining clinical risk factors** Figure 4(a) shows the distribution of different chronic conditions and risk factors in the entire population vs ventilated population. We observe that the admitted patients had high prevalence of hypertension, obesity, smoking, and diabetes (present in around 60-65% of the population), while hyperlipidemia and coronary artery disease (CAD) were observed in 40% and 23% of the cohort, respectively. However, the prevalence of all risk factors was substantially higher in the ventilated population. Diabetes, obesity, and hypertension were present in more than 90% of patients, establishing these as important risk factors for ventilation, while CAD and diabetes patients had the highest ratio of *ventilated/total population*.



**Figure 4.** (a) Risk factor prevalence in ventilated patients compared to the total population. (b) The distribution of model predicted risk scores for ventilation outcomes across the test cohort. The bars show the range while the mean score is showed as a line across the bar. All chronic conditions lead to a higher predicted risk while the mean scores were highest for patients with CAD and Diabetes (consistent with the ground truth trends observed for ventilated patients in (a)).

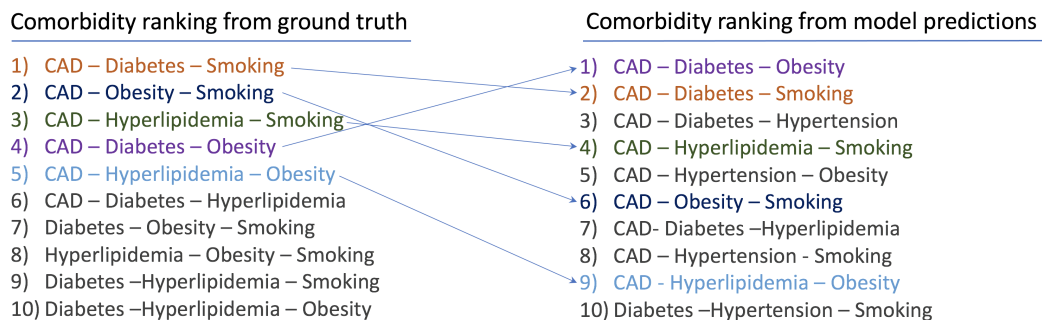
**Univariate analysis of model predicted risk scores** Figure 4(b) provides the model predicted ventilation risk score for patients with or without a risk factor. Patients with any of the risk factors were predicted to have a higher risk than patients without a risk factor, which is consistent with ground truth observation. The average predicted risk scores were highest for CAD and diabetes patients, with CAD patients having the largest variance. Between all risk factors, diabetes and hypertension patients had the highest increase in risk compared to the non-diabetic or non-hypertension patients.

We further studied the model predicted risk scores across different patient demographics in terms of age and sex (shown in Fig. 5). Patient race was excluded from our analysis due to insufficient coverage across different racial groups. Amongst male and female sex, the model predicted higher risk scores for male patients compared to their female counterparts, consistent with the ground truth for ventilation cases observed at 7.9% in male and 4.3% in female patients, respectively. For different age groups, the patients under 30 years of age were predicted to be at very low risk even in the presence of risk factors, again consistent with ventilation outcomes observed in the data (1/164 patients under 30). The model predicted risk increased for patients with age when they did not have any risk factor. However, in the presence of a chronic condition, the model gave a larger range of risk scores to CAD patients between ages 30-65, although the mean scores for other chronic conditions remained similar across ages 30-65 (ventilated = 51/883) and the 65+ age group (ventilated = 50/614). A detailed analysis of model performance for these sub-cohorts via AUROC measures is provided in Supplementary Table 2 and 3.



**Figure 5.** The influence of chronic risk factors across sex and age groups on the TRANSMED predicted risk scores for ventilation outcome ( $T_f = 3$ ).

**Multivariate analysis of model predicted risk scores** Finally, we introduce our methodology for evaluating the model predictions in terms of multi-way feature interactions. Details of the multi-comorbidity generation and ranking process are provided in Supplementary “Implementation of Multi-Comorbidity Ranking” section. We compare a ranking of the top multi-comorbidities from six clinical risk factors as determined by their prevalence in the ventilated sub-cohort with a ranking derived through model predicted scores (Figure 6). The model predicted risk ranking on the right closely agrees with the ground truth ranking of comorbidities for ventilated patients, with top-5 (out of 30) ground truth comorbidity interactions, in top-9 comorbidities identified by the model. Our analysis establishes that the TRANSMED learns reliable risk scores across salient clinical risk factors and captures multi-way feature interactions consistent with ground truth observed for ventilation outcomes.



**Figure 6.** Size-3 comorbidities by prevalence in ventilated patients vs model predicted risk score. The top 5 (out of 30) comorbidities in the ground truth were found within the top 9 risk scores predicted by the model.

## Discussion

Set against the backdrop of COVID-19, we reviewed a number of challenges faced by health systems worldwide to develop improved risk stratification tools for pandemic responses. We created a rigorous model evaluation framework dedicating significant effort to explore the best settings for the benchmark methods, exploring different pre-training strategies (pre-train on only SRD cohort, or a merged SRD and COVID-19 patient cohort) and evaluating impact of data source selection (see results for 12 different combinations in Supplementary Table 5. The evaluations were performed across the four prediction task settings of varying complexity. Our conclusions are drawn from the resultant space of models representing the widest possible variation of data and neural architectures for our study cohorts. The following are key observations that emerge from our study.

**Transfer learning from existing disease can be key to modeling emerging infectious (and rare) diseases.** Pre-training TRANSMED's transformer-based multi-modal patient encoder layers on the larger SRD cohort consistently provided performance boost ranging from 8-17% for our model performance. While our insight into the experiments with transfer learning primarily arose from clinical intuitions, we confirmed the "transfer potential" by inspecting the overlap in the vocabularies, as well as similarities in outcomes as detailed in the previous section. To our knowledge, no other work has demonstrated the ability to train an effective deep learning model for COVID-19 by training on other pre-existing patient cohorts.

**Building models with imperfect multi-modal data.** Much of the initial literature on predictive models for electronic healthcare records is overwhelmingly biased towards models with diagnostic codes and drug codes. Studies based on procedure codes or time-series measurements constitute a small fraction. TRANSMED's best performance was driven by a combination of procedure codes, numeric laboratory measurements, demographics and knowledge of risk factors extracted from clinical notes. Given the historical importance of diagnosis codes and drug codes in building predictive models and their relative under-performance in our study raises key questions about the utility of each data source. As Figure 1 illustrates, we found diagnostic codes are coded in sparsely. While medication data is available more continuously, it does not change frequently to suggest changes in a patient's condition. From this perspective, our best performing input combination is strongly intuitive. Observing critical procedures such as a radiological test or heparin therapy allows a model to escalate a patient's severity level, observing measurements such lymphocyte counts over time allows the model to reason about the trend of infection levels, and the knowledge of demographic information such as age, race, and sex coupled with prior knowledge of baseline risk factors such as diabetes, hypertension, and CAD can guide a model's association with other symptoms and outcomes.

**Guiding training data complexity.** We observe that in a setting with imperfect data, we need to explicitly reason about the discriminative value of each data source. We also ensured the availability of sufficient training data for each feature introduced in every data source. We only included 8 laboratory measurements out of 1496 unique laboratory measurement codes by considering the number of patients who had available data (we set 1200 patients as a minimum threshold) and the minimum number of days results were available for each patient (set to two). Ensuring high overlap between medical codes was critical to the benefit of transfer learning as well. Considering that procedure codes have an overlap of 63.2% across SRD and COVID-19 patient cohorts, as compared to 88.88% for diagnostic codes, 94.04% for drugs, and 90.57% for laboratory measurements, it is safe to say that the discriminative nature of procedure codes was a dominant factor over vocabulary overlap. Introduction of each feature affects the learning complexity by increasing the number of model parameters. Implicitly and intuitively, we sought to maximize the ratio of information entropy in our training data for each model parameter.

**Merit of hierarchical approach for learning from sparse data.** We conclude this discussion by noting that our approach outperformed others by explicitly recognizing the sparsity in the training data. Instead of learning the association of different medical codes or laboratory measurements at an entire patient stay level, we sought to learn the dependencies at finer granularity of time. However, we also used the demographics information and risk factors as static attributes associated with each time interval. Given that most of the multi-modal data streams occur sparsely and irregularly, this design decision reduced the complexity of learning the association between "everything" but provided less ambiguous input for each sample in the transfer learning step. In short, we ensured that the information flowed across modalities within a single time interval via the pre-training/transfer learning step, and then across time through the temporal layer during the fine-tuning step. Given that both TRANSMED and BEHRT were trained on identical transfer learning settings, the strong performance gap of TRANSMED over BEHRT demonstrates the merit of our hierarchical modeling approach.

## Summary and conclusions

Our work shows that a transfer-learning approach that learns from prior and related EHR databases is a promising way to build predictive models for diseases with limited data. A key conclusion from our study is that hierarchical learning, that first models the interaction between various medical concepts over shorter intervals, and then learns temporal dependencies is effective for transfer learning across diseases where patient conditions evolve at different time-scales. Our methodology demonstrates that a neural architecture that integrates both static (such as demographics and clinical risk factors) and dynamic information (such as temporal lab measurements) in a fashion that is robust to the sparsity and irregularity of multi-modal data sources

is likely to provide the best predictive model for complex outcomes. We also propose a method for multi-way comorbidity analysis that can be extended to include a richer set of phenotypes and evaluate a model's ability to capture complex interactions between them. TRANSMED's ability to improve the prediction accuracy on complex tasks such as predicting the likelihood of ventilation seven days into the future by an average of 17.5% on AUROC and 34.84% for F1 score demonstrates the promise of our method and motivates further investigation.

## Data and Code Availability

The data that support the findings of this study are available from STANford medicine Research data Repository(STARR) (<https://starr.stanford.edu>) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. The code will be made available upon request.

## References

1. Barda, N. et al. Developing a covid-19 mortality risk prediction model when individual-level data are not available. *Nat. communications* **11**, 1–9 (2020).
2. Yan, L. et al. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2**, 283–288, DOI: [10.1038/s42256-020-0180-7](https://doi.org/10.1038/s42256-020-0180-7) (2020).
3. Razavian, N. et al. A validated, real-time prediction model for favorable outcomes in hospitalized covid-19 patients. *NPJ digital medicine* **3**, 1–13 (2020).
4. Estiri, H. et al. Predicting covid-19 mortality with electronic medical records. *NPJ digital medicine* **4**, 1–10 (2021).
5. Kar, S. et al. Multivariable mortality risk prediction using machine learning for covid-19 patients at admission (aicovid). *Sci. reports* **11**, 1–11 (2021).
6. Yang, Z., Dehmer, M., Yli-Harja, O. & Emmert-Streib, F. Combining deep learning with token selection for patient phenotyping from electronic health records. *Sci. reports* **10**, 1–18 (2020).
7. Wang, Z. et al. Hospitalised covid-19 patients of the mount sinai health system: a retrospective observational study using the electronic medical records. *BMJ open* **10**, e040441 (2020).
8. Schwab, P. et al. Real-time prediction of covid-19 related mortality using electronic health records. *Nat. communications* **12**, 1–16 (2021).
9. Liang, W. et al. Early triage of critically ill covid-19 patients using deep learning. *Nat. communications* **11**, 1–7 (2020).
10. Krystal, J. H. Responding to the hidden pandemic for healthcare workers: stress. *Nat. medicine* **26**, 639–639 (2020).
11. Chen, Q. et al. Mental health care for medical staff in china during the covid-19 outbreak. *The Lancet Psychiatry* **7**, e15–e16 (2020).
12. Kang, L. et al. The mental health of medical workers in wuhan, china dealing with the 2019 novel coronavirus. *The Lancet Psychiatry* (2020).
13. Lassau, N. et al. Integrating deep learning ct-scan model, biological and clinical variables to predict severity of covid-19 patients. *Nat. communications* **12**, 1–11 (2021).
14. Mizrahi, B. et al. Longitudinal symptom dynamics of covid-19 infection. *Nat. communications* **11**, 1–10 (2020).
15. Gao, Y. et al. Machine learning based early warning system enables accurate mortality risk prediction for covid-19. *Nat. Commun.* **11**, 5033 (2020).
16. Carmichael, H. et al. Learning from past respiratory failure patients to triage covid-19 patient ventilator needs: A multi-institutional study. *J. biomedical informatics* **119**, 103802 (2021).
17. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, 301–318 (PMLR, 2016).
18. Choi, E., Bahadori, M. T., Song, L., Stewart, W. F. & Sun, J. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 787–795 (2017).
19. Choi, E., Xiao, C., Stewart, W. F. & Sun, J. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. *arXiv preprint arXiv:1810.09593* (2018).

20. Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine* **3**, 1–9 (2020).
21. Venugopalan, J., Tong, L., Hassanzadeh, H. R. & Wang, M. D. Multimodal deep learning models for early detection of alzheimer’s disease stage. *Sci. reports* **11**, 1–13 (2021).
22. Li, Y. et al. Inferring multimodal latent topics from electronic health records. *Nat. communications* **11**, 1–17 (2020).
23. El-Sappagh, S., Alonso, J. M., Islam, S. R., Sultan, A. M. & Kwak, K. S. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer’s disease. *Sci. reports* **11**, 1–26 (2021).
24. Landi, I. et al. Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ digital medicine* **3**, 1–11 (2020).
25. Harutyunyan, H., Khachatryan, H., Kale, D. C., Ver Steeg, G. & Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. data* **6**, 1–18 (2019).
26. Thiagarajan, J. J., Rajan, D., Katoch, S. & Spanias, A. Ddxnet: a deep learning model for automatic interpretation of electronic health records, electrocardiograms and electroencephalograms. *Sci. reports* **10**, 1–11 (2020).
27. Altschul, D. J. et al. A novel severity score to predict inpatient mortality in covid-19 patients. *Sci. reports* **10**, 1–8 (2020).
28. Gue, Y. X. et al. Development of a novel risk score to predict mortality in patients admitted to hospital with covid-19. *Sci. Reports* **10**, 1–8 (2020).
29. Xu, W. et al. Risk factors analysis of covid-19 patients with ards and prediction based on machine learning. *Sci. reports* **11**, 1–12 (2021).
30. Satici, C. et al. Performance of pneumonia severity index and curb-65 in predicting 30-day mortality in patients with covid-19. *Int. J. Infect. Dis.* **98**, 84–89 (2020).
31. Li, Y. et al. Behrt: transformer for electronic health records. *Sci. reports* **10**, 1–12 (2020).
32. 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* **4**, 1–13 (2021).
33. Burn, E. et al. Deep phenotyping of 34,128 adult patients hospitalised with covid-19 in an international network study. *Nat. Commun.* **11**, 5009 (2020).
34. Datta, S. et al. A new paradigm for accelerating clinical data science at stanford medicine. *CoRR* **abs/2003.10534** (2020). [2003.10534](https://arxiv.org/abs/2003.10534).
35. Kocaman, V. & Talby, D. Spark nlp: Natural language understanding at scale. *Softw. Impacts* **8**, 100058 (2021).
36. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
37. Wang, P., Agarwal, K., Ham, C., Choudhury, S. & Reddy, C. K. Self-supervised learning of contextual embeddings for link prediction in heterogeneous networks. In *Proceedings of The Web Conference 2021* (2021).
38. Chung, J., Gülçehre, Ç., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* **abs/1412.3555** (2014). [1412.3555](https://arxiv.org/abs/1412.3555).
39. Wang, X., He, X., Cao, Y., Liu, M. & Chua, T.-S. Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 950–958 (2019).

## Author contributions statement

S.Tamang, S.C., K.A., C.R. conceived the study, P.M. generated EHR datasets, P.M. and S.Tipirneni performed statistical analysis, K.A., S.C. designed and implemented TRANSMED, S.Tipirneni implemented all baselines, K.A., C.H., S.Tipirneni performed model optimizations and benchmarking, C.H., S.Tang, V.K. developed pipeline for processing clinical notes, S.Tamang, M.B., K.A., S.C. designed experiments for interpretability, C.R., O.G., R.R. analyzed the results. All authors reviewed the manuscript.

## Additional information

**Competing Interests:** The concepts presented in this paper on the hierarchical multi-modal model design are protected with a U.S. provisional patent 63/244,067. All authors have no conflicts to declare.



## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TransMEDsupplementary1008.pdf](#)