Continual Deterioration Prediction for Hospitalized COVID-19 Patients

JIACHENG LIU†, MEGHNA SINGH∗, CATHERINE ST.HILL†, VINO RAJ†, LISA KIRKLAND†, and JAIDEEP SRIVASTAVA∗

Leading up to August 2020, COVID-19 has spread to almost every country in the world, causing millions of infected and hundreds of thousands of deaths. In this paper, we first verify the assumption that clinical variables could have time-varying effects on COVID-19 outcomes. Then, we develop a temporal stratification approach to make daily predictions on patients’ outcome at the end of hospital stay. Training data is segmented by the remaining length of stay, which is a proxy for the patient’s overall condition. Based on this, a sequence of predictive models are built, one for each time segment. Thanks to the publicly shared data, we were able to build and evaluate prototype models. Preliminary experiments show 0.98 AUROC, 0.91 F1 score and 0.97 AUPR on continuous deterioration prediction, encouraging further development of the model as well as validations on different datasets. We also verify the key assumption which motivates our method. Clinical variables could have time-varying effects on COVID-19 outcomes. That is to say, the feature importance of a variable in the predictive model varies at different disease stages.

Key Words: COVID-19, Machine Learning, Patients Triaging, Patients Deterioration Prediction

1 INTRODUCTION

1.1 Background
The ongoing pandemic of COVID-19 has brought critical challenges to healthcare providers[COVID et al. 2020; Garg 2020; Wu et al. 2020a]. Many COVID-19 positive individuals deteriorate rapidly after the initial onset of symptoms and require hospitalization and intensive care including the use of mechanical ventilators and other intensive supportive measures for survival. There is an urgent need for effective COVID-19-specific care protocols or decision-aiding tools addressing several key clinical problems, such as predicting clinical deterioration and risk of mortality, estimating length of stay, and predicting usage of mechanical ventilators[Bai et al. 2020; Haimovich et al. 2020; Jiang et al. 2020; White and Lo 2020]. Ideally, given enough COVID-19 patients’ data, a data-driven approach would model these problems as either classification or regression tasks. However, this effort is largely impeded by the lack of privacy-preserved and law-compliant shared data.

1.2 Related Work
Machine learning techniques have been applied to aid triaging decisions for COVID-19 patients and have shown promising results. Both [Liang et al. 2020] and [Wu et al. 2020b] build logistic regression models to predict whether a patient will develop critical conditions such as admission to ICU, need of invasive ventilators or death. These models are based on clinical symptoms, lab results, radiology report, medical history and demographics. [Liang et al. 2020] report an AUROC of 0.88 on independent test sets. [Wu et al. 2020b] validate their model on 5 cohorts collected across hospitals in Belgium, China and Italy. The AUROC ranges from 0.84 to 0.89. However, these models are one-shot classification models working at the time of admission, rather than models making daily predictions. [Yan et al. 2020] released a publicly available COVID-19 electronic medical records dataset consisting of 485 COVID-19 patients which can be used for various research purposes. We will refer to this dataset as Wuhan Dataset in the rest of this paper, as all patient data is collected from Wuhan, China. In their work, XGBoost[Chen and Guestrin 2016] models are built to predict risk of mortality for patients’ last day records. The result shows 97% Accuracy on test and over 0.9 F1 score for both who survived and who died. However, these retrospective models are not useful in an operational setting because they rely on lab results on the day of the final outcome. We argue that their retrospective study cannot be directly applied to real-world COVID-19 deterioration prediction scenarios, for two reasons, (i) It is impossible to know a patient’s exact outcome date in advance. (ii) Variables may play time-varying roles during different stages of disease progression.

1.3 Contributions
Since “retrospective models” are not useful in an operational setting, we propose a temporally stratification approach to train “prospective models” that can make accurate predictions every day on the final outcome based on lab results, vitals, demographics and other medical data. The problem of predicting risk of mortality is known as the deterioration prediction problem. During the training phase of the proposed method, we stratify patients by their days until outcome, namely remaining length of stay. Secondly, a strata classifier is learned to predict the remaining length of stay for every patient. For each stratum, a risk of mortality model is built using all patient data that fall into this stratum. While testing and running, the predicted value is calculated as the dot product of strata probability distribution given by the strata classifier and the risk of mortality score given by each stratum model. Thus, our contributions are twofold.

(i) Verified the assumption that clinical variables play time-varying roles during different stages of disease progression.
(ii) Proposed a temporal stratified model to solve the issue of time-varying effects.

The rest of the paper is organized as follows. In section2, we first examine the original problem definition used in [Yan et al. 2020], then propose a new and more realistic definition of the deterioration prediction problem for COVID patients. Notations and abbreviations are also introduced there. In section3, we present empirical evidence which support the claim that clinical variables have time-varying effects. In section4, we elaborate our two-step stratification-prediction method and experiment results are highlighted in section5. Finally, drawbacks of this study and future work are discussed in section6.
The main goal of this preliminary study is to predict whether patients will survive at end of their hospitalizations. This decision-making tool may ultimately assist first line hospital workers in prediction of outcomes for hospitalized COVID-19 patients, either admitted to general wards or the intensive care unit (ICU). Other predictive goals, such as estimated hospital length of stay, are also crucial for hospital resources management during the COVID-19 pandemic.

We begin by introducing the notations, as shown in Table 1. As line 5 in Table 1 illustrates, the first subscription of a variable denotes time, namely days in the settings of this study on patients', while the second one denotes $i$ th patient. It is worth noting that for simplicity, we omit the subscription of patient index most of the time, as all the operations can be trivially applied to every patient, except when discussing evaluation metrics of model performance, where subscriptions of patient index are explicitly needed. We reiterated the problem definition in [Yan et al. 2020] using the notations in Table 1.

**Problem Statement 0** (Retrospective Modeling of COVID-19 Deterioration): Given only COVID-19 patients’ laboratory results as input data, we build regularized binary classifier $C$ to minimize the cross entropy between actual outcome $y_i$ and the predicted outcome $\hat{y}_{(t),i} = C(x_{(t),i})$, where $t(i) = n_{a,i}$.

In the definition above, $x_{(t),i}$ refers to the last available clinical data readings, or the closest data record to the discharged date, regardless of the method used to impute and process the data. There are three major limitations of this problem formulation.

- The entire dataset is underutilized. There are about 6,000 records in the raw dataset, however a majority of those are not used during model training since each patient will only have one record, the closest record to their discharged date, in the processed dataset.

- On the contrary to the author’s claim, this approach is operationally infeasible. During training, retrospective models are built upon patients’ last records in the hospital, but in the real world, we will never know in advance if a record is one’s last record. Therefore, the model is inapplicable due to a potential data distribution difference between training samples and real-world testing samples.

- It is reasonable for the model built and tested using only the most recent data record which is the closest to the discharged date, or, the last records, to have better performance than models built and tested otherwise, because of the continuity of time, samples tend to be more separable as they approach closer to their discharged date/death date.

These discussions around potential improvements on the original work inspire our definition of the deterioration problem as well as the proposed method to solve it. Instead of the retrospective version of problem definition stated in Problem Statement 0, we propose the predictive version of the COVID-19 deterioration problem.

**Problem Statement 1** (Predictive Modeling of COVID-19 Deterioration): Assuming there are $n_p$ hospitalized COVID-19 patients. For patient $i, 1 \leq i \leq n_p$, the daily feature vector is $x_{i,t}$. Admission and discharged datetimes are defined as $T_{a,i}$ and $T_{e,i}$. The goal is to build binary classifier $C$ which minimizes the loss function:

$\mathcal{L} = \Omega(C) + \sum_{i=1}^{n_p} \sum_{t=T_{a,i}}^{T_{e,i}} c(C(x_{i,t}), y_{t})$

where $\Omega(C)$ denotes the regularization loss and $c$ is the cross-entropy.

As we can see, instead of building models upon last records, the new problem statement will work for every patient every day. What’s more, for patients whose conditions are progressively worsening, in order to reflect the increasing importance of near-term predictions, we can assign weights $\lambda_t$ to days in the loss function.

---

1. In theory, the choice of prediction frequency or time granularity can be arbitrary. It can be hourly, daily or several times a day or whenever there is new patient data available. We choose to do daily predictions here in alignment with the original work.

2. Though not explicitly stated in the paper [Yan et al. 2020], this can be inferred from the data preprocessing and model training/evaluation codes released in the github repo: https://github.com/HAIRLAB/Pre_Surv_COVID_19
i.e., \( L = \Omega(C) + \sum_{i=1}^{n_{\text{p}}} \sum_{t_T_{i=1}}^{T_{i=1}} \lambda_t \text{ce}(C(x_{i1}), y_i) \) so that it will regularize the classifier toward more accurate near-term predictions.

3 TIME-VARYING EFFECTS OF CLINICAL VARIABLES

We motivate our methods by showing the evidence for the following assumption:

**Assumption 1**: Clinical variables have time-varying effects on COVID-19 deterioration outcomes.

Some features may be important for early stage COVID-19 patients and others may become important later as the disease progresses. To investigate the validity of the assumption, we will conduct the following experiments.

[Yan et al. 2020] found three key features, namely, lactic dehydrogenase (LDH), lymphocytes and high-sensitivity C-reactive protein (hs-CRP), as most important features in the model. Using the Wuhan Dataset, we built models for each date, where the input to the model was the latest laboratory result values of the three key features for patients currently admitted to the hospital. A supervised XGBoost classifier [Chen and Guestrin 2016] was used similar to the one used in [Yan et al. 2020]. Each model was trained on the training set data, and then we tested the performance of the model in predicting patient mortality on an external test set. Fig. 1 and 2 show the performance of the models for predicting survival and death respectively.

As can be seen from these figures, the baseline models using latest available laboratory results were able to predict patient survival with high accuracy within few days of being admitted to the hospital. But the models were random at best in predicting mortality.

Fig. 1. Baseline daily models using global time – Accuracy for Patient Survival

Fig. 2. Baseline daily models using global time – Accuracy for Patient Death

Moreover, Fig. 3, which shows feature importance for each of these models, strengthens our assumption that the importance of features may change with disease progression. Although further investigations with more data and hence, more statistical power, are needed to substantiate our findings, nevertheless, the analyses above clearly demonstrate the value and possibility to improve from the “one model for all” approach.

4 TEMPORAL STRATIFICATION

4.1 Patient Alignment

When patients are admitted to the hospital, they are at different stages in the course of COVID-19. What’s more, due to demographics, comorbidities and other risk factors, the rate of disease progression or recovery may vary from person to person. In a sense, this implies that the clinical data from these patients is heterogeneous. Moreover, as Assumption 1 states, clinical variables may play time-varying roles during the entire course of a disease. Thus, learning one model for all progression stages of a disease may not be the best solution. To tackle this issue, we may learn different models for each stage. First, patients’ time series data is aligned by a well-defined reference time point, such as days of symptom onset or days to outcome. Next, patient data can be either temporally binned, or stratified according to clinical definitions. The goal is to roughly segment one patient stay into separate stages. Finally, given enough data, predictive models are developed for each stratum.

Fig. 4 lists two common patient alignment methods, (b) aligned by the endpoint/discharged date, (c) aligned by the first day of hospitalization.

Fig. 4. Alignment methods
admission, as well as (a) the natural calendar time. Though aligning by symptoms onset or day of contact would be an ideal way, especially for an epidemiology study; for infectious diseases like COVID-19, during a global massive pandemic, it is not feasible to accurately trace and track the symptom onset date or exact contact date for every patient. Therefore, it is reasonable to choose patients’ discharged/outcome date as a reference time point. In this paper, patients are aligned by their outcome days and subsequently temporally stratified.

4.2 Stratification-Prediction Framework

In light of previous sections, the operationally feasible solution to Problem Statement 1 should also address Assumption 1 and preferably, be able to be incorporated into other important prediction goals such as length of stay. While the method in this section is developed for prediction of COVID-19 patients’ clinical deterioration while hospitalized, the general framework should also apply to other diseases, as long as the assumption still holds. The proposed method should work with both general wards and ICU patients, even though ICU data is expected to be denser and more varied. For simplicity and consistency with the original Wuhan Dataset, we won’t distinguish between general ward patients and ICU patients. We begin with discussions around the patient alignment problem, then move on to the proposed method, Temporally Stratified Model.

Fig. 5. Temporal stratification

Fig.5 shows that after aligning patients by discharged/outcome dates, we binned their time series data into predefined strata. The strata can be thought of as hidden states that control the distribution of clinical variables during a period of time. As a general framework, we can arbitrarily choose clinically meaningful strata definitions as long as the label would fit into a multi-class classification problem. In this pilot study, we stratified patients by their remaining length of stay. Therefore, the task of stratification is now effectively equivalent to the task of predicting the remaining length of stay. By right-aligning COVID-19 patients and temporally stratifying time series data, we decomposed the Problem Statement 1 into the following two sub-problems, Strata Assignment and Stratum-wise deterioration Prediction. Though there are algorithms which can implicitly model latent states, we chose to explicitly model the inferring of strata/hidden states as Strata Assignment.

As Fig.5 demonstrates, strata are mutually exclusive periods of time, yet collectively cover all days. Formally, let $S$ denote a set of time points delimiting neighboring stratum, then $S = \{d_1, d_2, \ldots, d_{n_s-1}\}$, where $0 \geq d_1 \geq d_2 \geq \cdots \geq d_{n_s-1}$. $|S| = n_s - 1$ since $n_s - 1$ points are needed to partition the timeline into $n_s$ segments. In Fig.4, $S = [-1, -2, -4, -7, -13]$. To be more specific, given $S$, strata are defined as the following half-open half-closed time periods, $s_1 = (-\infty, T_e + d_{n_s-1}]$, $s_{n_s-2} = (T_e + d_2, T_e + d_1]$, $s_{n_s-1} = (T_e + d_1, T_e]$. Notice that the data on the last day is discarded during training and testing since it would be too late for physicians to take any actions on the patients. Since all data are right-aligned, we can omit $T_e$ and then the notations would become $s_1 = (-\infty, d_{n_s-1}]$, $s_{n_s-2} = (d_2, d_1]$, $s_{n_s-1} = (d_1, 1)$.

Problem Statement 2 (Patient Strata Assignment) Assuming strata $S$, given daily clinical variables $x_t$, learn a multi-class stratification classifier $F$ which determines a point mass probability function over $n_s$ strata, namely $F(x_t) = \{p_1, p_2, \ldots, p_{n_s}\}$, the predicted probability of being in each stratum. We have $\sum_{k=1}^{n_s} p_k = 1$, $p_k \in [0, 1]$, for $0 \leq k \leq n_s$. Since, we designate the Strata Assignment task to be a daily task for every patient, the loss function can be written as:

$$L = \Omega(F) + \sum_{i=1}^{n_p} \sum_{t=1}^{T_e} ce(F(x_{t+1}), s_{t,i})$$

Problem Statement 3 (Stratum-wise Deterioration Prediction) Assume there is a solution $F$ to the Strata Assignment problem. For every stratum $k = [a, b] \in S$, for any patients who have data during this stratum defined period of time, learn a binary classifier $M^k$ that takes feature vectors and outputs a probability score $\hat{y}_t^k = P(r(y_t|s_t = k) = M^k(x_t))$, for $t \in [a, b]$. The loss function for stratum $k$ can be written as:

$$L^k = \Omega(M_k) + \sum_{\text{patients} \in S_k} \sum_{t = \max(T_e, b-1)}^{a} ce(M_k(x_t), y_t)$$

If there are $n_s$ strata, there will be $n_s$ stratum-wise deterioration prediction models. We now describe all the necessary steps of the proposed solution to Problem 1, 2 and 3. As illustrated by Fig.6, given any strata definition $S$, proceed with the following steps:

Fig. 6. Calculate risk scores
a. Temporally stratify patients into $n_s$ strata. The strata definition can be either given by medical experts or derived from the data.
b. Build strata classifier $F$ to assign stratum for every COVID-19 patient every day. If temporally stratified, it is equivalent to predict the remaining length of stay.
c. Build stratum-wise classifiers for every stratum $(a, b)$. A double truncation is implied here, meaning the training set for a particular stratum only contains those patients who survived at least $a$ days (left truncation) and had valid data during this stratum period(right truncation).
d. When testing, the outcome is given by the weighted average of each stratum-wise model, namely $\hat{y}_t = p_t \cdot M_t$

The risk of mortality prediction $\hat{y}_t$ is produced by the dot product of strata probability vector $p_t$ and the vector of predicted outcomes from $n_s$ stratum-wise models, $M_t = [M^1(x_t), M^2(x_t), \ldots, M^{n_s}(x_t)]$, where $M^k(x_t)$ is the prediction made by the $k$-th stratum risk prediction model. One issue with this approach is the choice and definitions of temporal strata, which affects the number of strata needed. There is no definite answer to these questions. To determine a proper definition, we can rely on medical knowledge, any reasonable heuristics, or even unsupervised time series clustering algorithms. As for the problem of determining the number of strata, we propose that given there is enough data within each stratum, having more strata than necessary will not significantly decrease the performance in terms of the total loss $\mathcal{L}$. In the worst case scenario, stratum-wise deterioration classifiers are similar to each other and is equivalent to the unstratified model trained using all data.

5 EXPERIMENTS

5.1 Wuhan Dataset

The *Wuhan Dataset* belongs to COVID-19 patients from the region of Wuhan, China and is from a mixed population of hospitalized patients in general wards (mild, severe) and ICU (critical). There are 375 patients in the training set, where 201 patients eventually recovered, while 174 died (46%). For each patient, input consists of values for 76 variables (lab results). Table 2 and Table 3 provide more details. An additional 110 patients (12% died) are in the test set, but that dataset has information for only 3 laboratory results.

Fig. 7 shows the right-aligned data in the *Wuhan Dataset*. Each patient had medical tests performed at different points in time during their hospital stay, and every day when such data was available is shown as a darker point based on time to outcome. All stays have been right aligned. Moreover, the length of stay is indicated by the darkness of color marked. For example, patient 252 stayed in the hospital for 32 days and had data available at days -31, -29, -27, -23 etc. The outcome for this patient was death. As another example, patient 54 stayed in the hospital for 7 days, had data available on -6, -3 and 0 days and had an outcome of survival. Another observation from this figure is that patients who did not survive had shorter stays in the hospital when compared to those who survived, which can be seen from the darker areas in the top half of the figure. Notice that the density of tests is much closer to the right side, which illustrates the need to have a more accurate estimate of the patient’s condition closer to the end of stay. This intuition is the basis for having time strata of differing lengths as shown in Fig. 5.

5.2 Results and Analysis

We built supervised classifiers using an XGBoost algorithm ([Chen and Guestrin 2016](#)), which is a tree-based gradient boosting algorithm. Models were built for each stratum, $S = -1, -2, -4, -7, -13$, which takes as input, all data that was associated with days within the stratum. For example, the model for stratum -1 would get trained on patient data in days range $(-2, -1]$. Similarly, the model for stratum -13 would use all data from date of admission up to 13 days before the outcome, i.e. $(−∞, −13]$ for training. Prediction by model $M^k(x_t)$ in the $k$-th stratum is a probability of death within $k$ days.

Various configurations of XGBoost were used to train the models, and the configurations with the best outcomes on the validation set were used to make predictions on the test set. Here, we predefined the strata definition to strike a balance between having enough number of available patients in each strata and modeling the underlying disease progression stages pointed by medical literature ([Bellino et al. 2020](#); [Garg 2020](#); [Polak et al. 2020](#); [Wu et al. 2020](#)).

Compared to the baseline models in Fig. 1, Fig. 2, and Fig. 3, the stratum-wise models (results shown in Fig. 8, Fig. 9, Fig. 10, and Table 4) perform consistently for death outcome, with precision close
Table 2. **Wuhan Dataset, Training set**

|        | Critical | Severe | Mild | Total     |
|--------|----------|--------|------|-----------|
| Recovered | 3 (0.8%) | 13 (3.47%) | 185 (49.33%) | 201 (53.60%) |
| Died    | 148 (39.47%) | 14 (3.73%) | 12 (3.2%) | 174 (46.4%) |
| Total   | 151 (40.27%) | 27 (7.20%) | 197 (52.53%) | 375 (100.00%) |

Table 3. **Wuhan Dataset, Test set**

|        | Critical | Severe | Mild | Total     |
|--------|----------|--------|------|-----------|
| Recovered | 13 (11.82%) | 34 (30.91%) | 50 (45.45%) | 97 (88.18%) |
| Died    | 9 (8.18%) | 4 (3.64%) | 0 (0.00%) | 13 (11.82%) |
| Total   | 22 (20%) | 38 (34.55%) | 50 (45.45%) | 110 (100.00%) |

**DISCUSSIONS**

One benefit of the temporal stratification approach is that it incorporates the change of feature importance across different stages. Another advantage is that it allows a multi-task learning of outcome predictions and resources usage predictions, since the stratum contains information on how long patients will continue to stay. In the temporal stratification approach, the remaining length of stay is

6 to 1 about 2 weeks before the outcome and sensitivity close to 1 about 1 week before the outcome. All experiments are coded in python. We use XGBoost for both stratum classifiers and stratum-wise prediction models. The stratum-wise model use decision trees as base classifiers. Hyper-parameters are tuned based on a 20% split(validation set) within the training set. The maximum depth of each base classifier is set to be 4. Both stratum classifiers and stratum-wise prediction models are trained with a learning rate of 0.2 and a subsample rate of 0.9. No L1 or L2 regularization for stratum-wise classifiers. We set a 0.02 L2 regularization for strata classifier.

To evaluate the performance of the combined models (strata assignment and stratum-wise prediction), we split the dataset twice to generate training cohort (60% of the patients), validation cohort (20% of the patients) and the test cohort (20% of the patients). In order to estimate the standard error, the stratified group 5-fold cross validation is repeated 100 times. The AUROC is 0.9805 with a standard error of 0.0198. More details are shown in Table 5.

We also tried to use the mean, slope, standard deviations and the first order difference of lab results. However, due to the sparseness of the data, additional features show little or no improvements. For a particular lab test, during the entire stay, one patient may only take 2-3 times. To some extent, the sparseness of lab results prevents the usage of more complicated sequential models like hidden Markov models[Rabiner and Juang 1986] and recurrent neural networks[Hochreiter and Schmidhuber 1997].
Continual Deterioration Prediction for Hospitalized COVID-19 Patients

Table 4. Performance of the Stratum-wise Models

| Stratum | #Patients | # Deaths | \( y = 1 \) means survived | \( y = 1 \) means died |
|---------|-----------|----------|-----------------------------|-------------------------|
| \((-\infty, -13]\) | 46 | 7 | 0.897 0.981 0.912 | 0.781 0.684 0.722 0.650 |
| \((-13, -7]\) | 99 | 13 | 0.963 0.938 0.991 | 0.914 0.900 0.973 0.837 |
| \((-7, -4]\) | 48 | 12 | 0.960 1.000 0.923 | 0.962 0.952 0.909 1.000 |
| \((-4, -2]\) | 40 | 12 | 0.983 1.000 0.967 | 0.983 0.979 0.958 1.000 |
| \((-2, 1]\) | 34 | 12 | 0.957 0.917 1.000 | 0.917 0.909 1.000 0.833 |

Table 5. Performance of the Combined Models

| AUC ROC(se) | AUC PR(se) | Accuracy(se) | F1 score(se) | Precision(se) | Recall(se) |
|-------------|------------|--------------|--------------|---------------|------------|
| 0.9805      | 0.9720     | 0.9353       | 0.9134       | 0.9186        | 0.9105     |
| (0.0198)    | (0.0142)   | (0.0225)     | (0.0344)     | (0.0469)      | (0.0479)   |

- Implicitly modeled and predicted by the stratum classifier. Problem 2 is equivalent to predict the remaining length of stay. Still, the definition of strata remains somewhat arbitrary and is yet to be explored. Besides temporal stratification, we can also use predefined medical phenotypes, or unsupervised clustering algorithms to identify latent states. Furthermore, we can impose restrictions on transitions between strata based on medical professional’s advice. The idea is to learn the state-space model which governs the progression of diseases.

There are also several limitations of this study. First of all, the method for deriving strata definition is somewhat heuristic. One can have more or less strata, based on the size of the dataset and the nature of the disease. An ideal strata definition should satisfy two criterion: (i) each stratum should have enough data and patients, so that k-fold cross validation can be conducted during model training. (ii) Breakdown points should be carefully determined in order to approximate the actual disease progression trajectory. Secondly, there are limitations of the dataset, namely the test set does not have as many variables as the training set does. Therefore, we are not able to utilize the full 100 variables on both training and testing sets. What’s more, the dataset only contains sparse lab variables, while medical history, demographics and vitals are not published. For a particular lab test, during the entire stay, one patient may only take 2-3 times. To some extent, the sparseness of lab results prevents the usage of more complicated time-dependent models. Last but not least, variations in patients and disease severity groups are not taken into account. Comorbidities, ages, and gender are often confounders which affects the final outcome and also importantly, the speed of disease progression or recovery. The stratification-prediction approach is assuming all the patients who finally died of COVID are approximately following the same progression trajectory, of which the variation is controlled by the width of each stratum. However, this may not hold in the real case or in other diseases.

For future work, we consider using sequential models and applications to other diseases and prediction targets, such as ICU admissions and the use of ventilators. We list possible extensions here.
7 CONCLUSIONS

In this pilot study, we verify the assumption that clinical variables are likely to have time-varying effects on COVID-19 deterioration outcomes. To address this issue, we propose temporally stratified models, and effectively delineate the deterioration prediction problem into two sub-problems: strata assignment and stratum-wise risk of mortality prediction. The method can be generalized as a stratification-prediction approach framework. Prototype models are developed using the Wuhan Dataset. Our experiments show promising results, encouraging further investigations. We are currently collecting data for COVID-19 hospitalized patients admitted into a large hospital system in an IRB–approved study. In the future, we plan to apply the proposed method to other disease and prediction targets, derive models using this hospital based dataset, and validate models in a future prospective study.

ACKNOWLEDGMENTS

We would like to thank the authors of Wuhan Dataset [Yan et al. 2020] for their decision to make the dataset publicly available.

REFERENCES

Xiang Bai, Cong Fang, Yu Zhou, Song Bai, Zaiyi Liu, Liming Xia, Qianlan Chen, Yongchao Xu, Tian Xia, Shi Gong, et al. 2020. Predicting COVID-19 malignant progression with AI techniques. (2020).

Stefania Bellino, Ornella Punzo, Maria Cristina Rota, Martina Del Manso, Alberto Matteo Urdiales, Xanthi Andrianou, Flavia Riccardo, et al. 2020. COVID-19 disease severity risk factors for pediatric patients in Italy. Pediatrics 146, 4 (2020).

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 785–794.

IHME COVID, Christopher JL Murray, et al. 2020. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. MedRxiv (2020).

Shikha Garg. 2020. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—COVID-NET, 14 States, March 1–30, 2020. MMWR. Morbidity and mortality weekly report 69 (2020).

Adrian D Haimovich, Neal G Ravindra, Stoytcho Stoytchev, H Patrick Young, Francis P Wilson, David van Dijk, Wade L Schulz, and R Andrew Taylor. 2020. Development and validation of the quick COVID-19 severity index: a prognostic tool for early clinical decompensation. Annals of emergency medicine 76, 4 (2020), 442–453.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.