DistCare: Distilling Knowledge from Publicly Available Online EMR Data to Emerging Epidemic for Prognosis

Liantao Ma¹,³, Xinyu Ma¹,³, Junyi Gao¹, Xianfeng Jiao¹, Zhihao Yu¹, Chaohu Zhang¹,³, Wenjie Ruan⁴, Yasha Wang⁵,², Wen Tang⁵, Jiantao Wang⁶

¹Key Laboratory of High Confidence Software Technologies, Ministry of Education, Beijing, CN
²National Engineering Research Center of Software Engineering, Peking University, Beijing, CN
³School of Electronics Engineering and Computer Science, Peking University, Beijing, CN
⁴Department of Computer Science, CEMPS, University of Exeter, UK
⁵Division of Nephrology, Peking University Third Hospital, Beijing, CN
⁶Center for Intelligent Healthcare, Coventry University, UK

{malt,maxinyu,wangyasha}@pku.edu.cn

ABSTRACT
Due to the characteristics of COVID-19, the epidemic develops rapidly and overwhelms health service systems worldwide. Many patients suffer from life-threatening systemic problems and need to be carefully monitored in ICUs. An intelligent prognosis can help physicians take an early intervention, prevent adverse outcomes, and optimize the medical resource allocation, which is urgently needed, especially in this ongoing global pandemic crisis. However, in the early stage of the epidemic outbreak, the data available for analysis is limited due to the lack of effective diagnostic mechanisms, the rarity of the cases, and privacy concerns. In this paper, we propose a distilled transfer learning framework, DistCare, which leverages the existing publicly available online Electronic Medical Records to enhance the prognosis for inpatients with emerging infectious diseases. It learns to embed the COVID-19-related medical features based on massive existing EMR data. The transferred parameters are further trained to imitate the teacher model’s representation based on distillation, which embeds the health status more comprehensively on the source dataset. We conduct Length-of-Stay prediction experiments for patients in ICUs on real-world COVID-19 datasets. The experiment results indicate that our proposed model consistently outperforms competitive baseline methods. In order to further verify the scalability of DistCare to deal with different clinical tasks on different EMR datasets, we conduct an additional mortality prediction experiment on End-Stage Renal Disease datasets. The extensive experiments demonstrate that DistCare can benefit the prognosis for emerging pandemics and other diseases with limited EMR.

CCS CONCEPTS
• Applied computing → Health informatics.

KEYWORDS
Electronic Medical Record, Healthcare Informatics, Prognosis, Transfer Learning

ACM Reference Format:
Liantao Ma¹,³, Xinyu Ma¹,³, Junyi Gao¹, Xianfeng Jiao¹, Zhihao Yu¹, Chaohu Zhang¹,³, Wenjie Ruan⁴, Yasha Wang⁵,², Wen Tang⁵, Jiantao Wang⁶. 2021. DistCare: Distilling Knowledge from Publicly Available Online EMR Data to Emerging Epidemic for Prognosis. In Proceedings of the Web Conference 2021 (WWW ’21), April 19–23, 2021, Ljubljana, Slovenia. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442381.3449855

1 INTRODUCTION
Since January 2020, the whole world has been facing an unprecedented pandemic crisis brought by COVID-19. The exponential growth of COVID-19 patients has brought massive pressure on the health systems tragically, such as overwhelming the national health service and exhausting the Intensive Care Units (ICUs). It is crucially essential to personalize prognosis for the individual patient by considering her/his specific health condition to enable a timely and early medical intervention. The accurate evaluation of inpatients’ health status is also critical for scheduling and optimizing limited hospital resources [16]. However, it is difficult for human physicians to evaluate patients’ health comprehensively and accurately identify the key factors, especially in the early stage of Emerging Infectious Diseases (EIDs) when the deterioration is usually not evident [24]. The precise risk prediction requires substantial clinical expertise and possibly years of experience [32]. For most EIDs (e.g., COVID-19, SARS), the prognosis performed by human physicians may not meet huge clinical demand, especially in developing countries, while clinical experience accumulation is time-consuming and challenging in the early outbreak of EIDs. At the early stage of EIDs, human physicians are lack of knowledge and experience for the diseases. So during COVID-19 treatment, physicians may omit certain ominous signs and miss the chance of early intervention, especially when the clinical resources are insufficient.

As a result, intelligent prognosis is in urgent need against EID and rare diseases. It not only assists physicians to perform early diagnosis, selects personalized treatments and prevents adverse outcomes, but also optimizes the allocation of medical resources.
and reduces the medical cost [43]. Recently, many deep-learning-based models have been developed to enable intelligent prognosis by analyzing Electronic Medical Records (EMR), including mortality prediction [28, 29], disease diagnosis prediction [12, 26], and patient phenotype identification [1]. To enrich the feature extraction and health status representation, most existing research works utilize sophisticated modules to extract health status representations that require a large amount of labeled training data.

However, existing AI-assisted health data analytic models and systems cannot be directly applied in the scenario of emerging epidemics, especially in the early stage of the epidemic, when there are only very limited medical data available for study. For example, by Jan 2, 2020, only 41 admitted hospital patients had been identified as having COVID-19 infection, which means that the meaningful data for physicians to study COVID-19 are highly insufficient [19]. Moreover, existing models require a large amount of data for training, but the quantity of labeled clinical data available for prognosis are insufficient as well in practice in the early stage of the EID outbreak [19]. One reason is that the precise diagnostic mechanism has not been established in the early outbreak. For example, before introducing the nucleic acid detection mechanism, it is difficult to confirm whether a patient is infected with COVID-19, so that researchers cannot acquire enough labeled data [19]. The privacy concern is another crucial reason to hinder the access of labeled clinical data. For example, at the beginning of the COVID-19 outbreak, sharing EMR data across different countries cannot be established timely due to the privacy and ethical consideration [20]. Thus the scarcity of labeled clinical data will decrease deep learning performance for the EID-related applications due to the potential over-fitting.

Recently some researchers try to make full use of the existing time series data through transfer learning to deal with clinical data scarcity. These transferring-based models are either focusing on transferring pre-trained models for a specific disease, or transferring general-purpose time-series features. 1) For instance, Doctor AI [3] and Gupta [14] train deep learning models at one hospital and transfer them to another hospital. These methods can only be applied to the same task with the same clinical features between the source and target dataset, while clinical features are usually not exactly the same in clinical practice. However, the extracted general-purpose features are also not suitable for specific clinical tasks. In the worst case, this can lead to negative transfer and model’s under-performance [38].

Therefore, for the prognosis of EIDs with limited data, such a research challenge remains: How to make full use of the existing publicly available EMR data to learn the robust health status representation when tackling different tasks with different clinical feature sets? In this paper, we propose a novel distilled transfer learning framework, DistCare, to distill knowledge from existing EMR data (i.e., source dataset) to the new dataset (i.e., target dataset). In summary, DistCare contributes to the community from the following aspects:

- We propose a medical feature embedding approach based on distilled transfer learning, DistCare, to perform clinical prediction for EIDs with limited data. In order to explore and leverage the features information that is only stored in the source dataset, the pre-trained model with all source features is treated as a teacher network to guide shared features’ embedding behavior.
- We conduct Length-of-Stay (LOS) prediction experiments for inpatients with COVID-19. The results show that DistCare consistently outperforms the baseline approaches under all evaluation metrics, especially when tackling insufficient data settings. Beyond COVID-19, in order to verify the applicability of DistCare when distilling the knowledge to perform different clinical tasks on different datasets, we also conduct mortality risk prediction for outpatients with End-Stage Renal Diseases (ESRD). The extensive experiments demonstrate that DistCare can significantly benefit the prognosis for pandemics and other diseases with limited EMR.
- As a proof-of-concept to demonstrate that DistCare can assist the prognosis, we also build a visualization system that can reveal the patient’s health trajectory for the prognosis. We release our code and the system at https://github.com/Accountable-Machine-Intelligence/DistCare.

2 RELATED WORK

2.1 Prognosis for COVID-19

Outbreaks of the COVID-19 epidemic have been causing worldwide health concerns and was officially declared a pandemic by the World Health Organization (WHO) on March 11, 2020. Although the ultimate impact of COVID-19 is uncertain, it has significantly overwhelmed health care infrastructure. All emerging viral pandemics can place extraordinary and sustained demands on public health systems and essential community service providers [31]. Limited healthcare resource availability will increase the chance of being infected while waiting for treatment and mortality rates [22]. This eventually leads to an increase of the severity of the pandemic. The rapidly growing imbalance between supply and demand for medical resources in many countries presents an inherent normative question: How can we make early and accurate risk prediction to allocate medical resources during a pandemic effectively?

Massive COVID-related research works focus on the severity of disease rather than the clinical outcome of mortality [8, 11, 42]. These studies answer critical clinical questions on COVID-19 evolution and outcomes, as well as potential risk factors leading to hospital and ICU admission. However, they cannot make individualized risk predictions for patients. Recently, Li et al., [44] use machine learning-based methods such as decision trees to make risk prediction for COVID-19 patients. Effective and reliable early risk prediction is still a crucial and urgent problem to optimize patient care and appropriately deploy healthcare resources during this pandemic.

2.2 Deep-Learning-Based EMR Analysis

With the prevalence of electronic healthcare information systems in various healthcare institutions, a large amount of Electronic Medical Records (EMR) have been accumulated over time [25, 36]. EMR is a type of multivariate time series data that records patients’ visits in
hospitals (e.g., diagnoses, lab tests). This provides essential healthcare information for the data-driven clinical status prediction [10]. Deep learning-based models have shown the capability to perform mortality prediction [9, 13, 17, 28, 40], patients subtyping [1], and diagnosis prediction [1, 5, 26, 30, 33, 35]. For most research, extracting advanced clinical features and learning the sparse EMR data’s compressed representation are fundamental procedures of clinical healthcare prediction.

EMR is longitudinally complex [6, 45]. Extracting the advanced clinical representation would introduce more parameters into the model, making the model more complex and challenging to train. For EIDs and some rare diseases, the quantity of labeled data is insufficient, which can not support a model to be trained thoroughly. In order to deal with this issue, some researchers try to introduce additional information about the data.

For example, GRAM [4] and KAME [27] incorporate the external medical information (e.g., ontologies of the medical codes), making the model trained more sufficiently. They exploit medical knowledge in the whole prediction process by using a given medical ontology (i.e., knowledge graph), such as the International Classification of Diseases (ICD), to learn the representations of medical codes and obtain the embeddings of medical codes’ ancestors. MIME [6] learns the multi-level embedding of data according to the knowledge about the inherent EMR structure (e.g., the multi-level relationship among medical codes). However, such external structured information and the extra knowledge about the data are often not easy to be accessed or used in the clinical practice for EIDs. Ontology information is usually designed to handle the medical codes. It is not suitable for dealing with numerical lab tests, which also are essential clinical features to capture health status.

On the other hand, some researchers try to explore the existing EMR data. Choi [3] empirically confirms that RNN models possess great potential for transfer learning across different medical institutions. Gupta [14] trains a deep RNN to identify several patient phenotypes on time series from MIMIC-III dataset, and then uses the features extracted by the RNN to build classifiers for identifying previously unseen phenotypes. However, these methods can only be utilized for the same tasks with the same clinical feature sets between source and target datasets. TimeNet [15] is pre-trained on non-medical time series in an unsupervised manner and further utilized to extract features for clinical prediction. Nevertheless, the extracted general-purpose features may not be suitable for exploring the specific clinical task, leading to negative transfer and limited performance.

### 3 PROBLEM FORMULATION

Many patients suffering from COVID-19 face severe life threats and need careful health monitoring in ICU. Besides, due to the newly emerging pandemic characteristics, a large number of patients need treatment during peak illness periods, which causes clinics and hospitals to be overwhelmed. Predicting remaining time spent in ICU (i.e., \textit{Length-of-Stay, LOS}) for admission can help assess the severity of illness and determine the value of novel treatments, interventions, as well as health care policies [34]. Moreover, it is also vital for scheduling and hospital resource management. Here we formally define our research problem below and provide the list of notations used in DistCare in Table 1.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_{\text{tar}}$</td>
<td>Groundtruth Label of LOS prediction at $t$-th record on target dataset</td>
</tr>
<tr>
<td>$\hat{y}_{\text{tar}}$</td>
<td>Prediction result at $t$-th record on target dataset</td>
</tr>
<tr>
<td>$y_{\text{src}}$</td>
<td>Groundtruth Label of prediction at $t$-th record on source dataset</td>
</tr>
<tr>
<td>$\hat{y}_{\text{src}}$</td>
<td>Prediction result at $T$-th record on source dataset</td>
</tr>
<tr>
<td>$\mathcal{R}_{\text{src}}$, $\mathcal{N}_i$</td>
<td>The whole source dataset with $N_i$ features</td>
</tr>
<tr>
<td>$\mathcal{R}_{\text{tar}}$, $\mathcal{N}_i$</td>
<td>The whole target dataset with $N_i$ features</td>
</tr>
<tr>
<td>$\mathcal{R}_{\text{src},\mathcal{N}_i}$</td>
<td>Source dataset (Only consists of $N_i$ features shared with the target dataset)</td>
</tr>
<tr>
<td>$r_i$</td>
<td>A time-series record of the $i$-th medical feature</td>
</tr>
<tr>
<td>$f_i$</td>
<td>Embedding of the $i$-th medical feature</td>
</tr>
<tr>
<td>$f'_i$</td>
<td>Re-encoded embedding of the $i$-th medical feature</td>
</tr>
<tr>
<td>$s$</td>
<td>Overall representation of patient’s health status</td>
</tr>
<tr>
<td>$\text{demo}$</td>
<td>The static baseline demographic information of the patient</td>
</tr>
<tr>
<td>$X_{\text{src}}$</td>
<td>Model/Embeddings/Parameters used in Source-Teacher model</td>
</tr>
<tr>
<td>$X_{\text{stu}}$</td>
<td>Model/Embeddings/Parameters used in Source-Student model</td>
</tr>
<tr>
<td>$X_{\text{tar}}$</td>
<td>Model/Embeddings/Parameters used in Target model</td>
</tr>
</tbody>
</table>

**Electronic Medical Records:** EMR is routinely collected patient observations from hospitals through the clinical admissions, including discrete time-series data (e.g., medication, diagnosis), continuous multivariate data (e.g., vital signs, laboratory measurements), and static baseline information (e.g., age, gender, primary disease). The static feature is denoted as $\text{demo}$. The admissions generating $N$ features such as different lab test results denoted as $r_i \in \mathbb{R}$. Each medical feature contains $T$ steps. As a result, such a clinical sequence can be formulated as a “longitudinal patient matrix” $r$, where one dimension represents medical features, and the other denotes record timestamps [25].

**Problem: Length-of-Stay prediction.** The prediction problem in this paper can be formulated as follows. Given historic EMR data of a patient, i.e., $(r_1, \ldots, r_N, \text{demo})$, the problem is to evaluate the patient’s health status at each clinical record and predict the remaining time spent in ICU, which is framed as a regression task. There are significant differences in health status among patients with the same length-of-stay but different outcomes. For instance, for patients being discharged in a few days, their health status should be much better than other patients, especially those who are dying in several days unfortunately. Concretely, we take the remaining days $t$ in ICU as the ground truth LOS label of survived patients’ records, and take the \textit{LIMIT} $- t$ as the label of deceased patients’ records. According to the statistics data, most inpatients with COVID-19 in ICU have been discharged within 35 days. As a result, we set the \textit{LIMIT} value as $2 \times 35 = 70$ in this work.

Specifically, the LOS to be predicted (i.e., $y$) for each patient’s records is defined according to the patient’s remaining time spent in ICU (i.e., $t$ days) and their outcomes.

$$
y = \begin{cases} 
\min(35, t), & \text{if discharged from ICU} \\
70 - \min(35, t), & \text{if died in ICU}
\end{cases}
$$

The predicted LOS $\hat{y}$ can be supposed as a health risk score. Patients with high-risk scores are facing a high probability of adverse outcomes and need emergency treatment. On the contrary, those with low-risk scores are in relatively stable health conditions. The predicted LOS health risk score indicates the remaining days to discharge when $\hat{y} < 35$. And $70 - \hat{y}$ indicates the remaining days to mortality when $\hat{y} > 35$. 

---

**Table 1: Notations Used in DistCare**
4 METHODOLOGY

4.1 Overview
DistCare learns to effectively embed the clinical time series based on a massive existing EMR source dataset. Such a transfer learning mechanism reduces the demand for training data on the target. The patient’s health status representation learning process is further guided by distilled transfer learning between the source dataset and target dataset. Figure 1 shows the framework of the proposed DistCare, which contains two key steps:

- **Teacher model’s healthcare representation learning:** Multivariate time series with all features are fed into the healthcare representation learning module as a teacher model to build the health status embedding in the source dataset.
- **Distilled transfer learning from student model to target model:** The student model on source dataset learns to embed the proper health status based on features shared with the target dataset, by imitating the teacher model’s embedding behavior. The pre-trained parameters of feature embeddings are transferred to the healthcare representation learning model on the target dataset, and further fine-tuned to perform the task-specific prediction.

4.2 Healthcare Representation Learning
We employ the basic patient health context embedding module inspired by ConCare [29]. There are three layers designed in this module, namely, the feature extraction layer, the self-attention layer, and the prediction layer. We utilize the multichannel GRU in the feature extraction layer to capture each medical feature’s patterns individually. Specifically, we apply $N$ different GRUs to embed the $N$ dynamic features. Each dynamic feature $f_i$ can be described as a time series $r_i = (r_{i1}, r_{i2}, \ldots, r_{iT})$, and will be fed into the corresponding GRU$_i$ to generate feature embedding:

$$f_i = \text{GRU}_i(r_{i1}, r_{i2}, \ldots, r_{iT}).$$  

(1)

And the static baseline demographic feature demo is mapped to the embedding $f_0$ with a full connection network: $f_0 = \text{demo} \cdot W_{\text{demo}} + b_{\text{demo}}$. The embeddings $f_i$ are stacked to generate the feature embedding matrix $F = (f_0, f_1, f_2, \ldots, f_N)^T$.

The self-attention mechanism is utilized to obtain information from the health context and capture correlations between medical features. This mechanism makes each feature adaptively interact with all other features. The re-encoded embeddings of heterogeneous clinical features are guaranteed to be in the same high-level feature space by the self-attention mechanism. Mathematically, the self-attention weight matrix of head $i$ is

$$A_i = \text{Softmax}(\frac{Q_iK_i^T}{\sqrt{d_k}}),$$  

(2)

where $Q_i = F \cdot W_i^Q$, $K_i = F \cdot W^K$, and $d_k$ is the size of the row vector of matrix $K_i$. And the result of feature interaction in head $i$ is calculated as:

$$\text{head}_i = A_iV_i$$  

(3)

where $V_i = F \cdot W_i^Y$. And finally, the embedding matrix $F^*$ after feature interaction is calculated as:

$$F^* = (f_0^*, f_1^*, \ldots, f_N^*)^T = (\text{head}_0 \oplus \text{head}_1 \oplus \ldots \oplus \text{head}_m)W^O$$  

(4)

Embedding of all features $f_i^*$ are integrated into an overall representation of patient $s$. The importance of medical features is interpreted by attention mechanism.

$$s = \sum_{i=1}^{N} \alpha_if_i^*$$  

(5)
where $a_i$ is the attention weight of each feature embedding $f_i$, generated by mean query feature embedding $q_{\text{mean}}$ and key feature embeddings $k_i$.

$$a_0, a_1, \ldots, a_N = \text{Softmax}(\zeta_0, \zeta_1, \ldots, \zeta_N)$$  

(6)

$$\zeta_i = \tanh(q_{\text{mean}} \cdot k_i^T)$$  

(7)

$$q_{\text{mean}} = \frac{1}{N+1} \sum_{i=0}^{N} f_i \cdot W_q, \quad k_i = f_i \cdot W_i$$  

(8)

Eventually, in the prediction layer, we apply a full-connection layer to predict the clinical task $y_T \in \mathbb{R}$.

$$y_T = s \cdot W_s + b_s$$  

(9)

and we adopt Mean Square Error (MSE) as the loss term $L_{\text{pred}}$:

$$L_{\text{pred}} = \text{MSE}(y_T, y_T^*) = \frac{1}{N} \sum_{i=1}^{N} (y_T^*(i) - y_T(i))^2$$  

(10)

Alternatively for predicting a binary classification label $y_T$, we apply the Sigmoid activation to Eq.9, and the Cross-Entropy loss will be applied for computing the prediction loss $L_{\text{pred}}$.

4.3 Distilled Transfer Learning

Based on the patient health status embedding module introduced above, we conduct feature-specific transfer learning on the feature extraction layer, since this layer mainly captures the general pattern of medical features, which is independent of patient cohorts and prediction tasks. Concretely, we transfer GRUs of shared features from the source model to the target model, to make up for the shortcomings of small data volume by obtaining knowledge from a larger existing dataset.

However, the source dataset’s useful information has not been sufficiently extracted, since private features in the source dataset remain unused. They can help capture correlations between features more sufficiently, thus generating a more comprehensive health status representation.

Therefore, we propose a distillation mechanism to construct a more reasonable source to transfer. Concretely, we divide the source model into two parts, the teacher model and the student model. The student model is trained on the source dataset with only shared features ($R_{\text{src}}$) and prepared to be transferred to the target model. While the teacher model is trained on the complete dataset with all features ($R_{\text{src}}$), and serves as an auxiliary representation extractor. The distillation mechanism aims to distill the teacher model’s knowledge to assist training on the student model, guiding the student to imitate the teacher’s behavior to obtain a more comprehensive representation of patients.

Specifically, the representation $s_{\text{stu}}$ generated by the student model should be able to imitate $s_{\text{tea}}$ generated by the teacher model as much as possible, by using a linear layer to transform the feature space. And the distillation loss ($L_{\text{dist}}$) is defined as the similarity of the two representations, which is calculated using KL-Divergence.

$$s_{\text{tea}} = s_{\text{stu}} \cdot W_{\text{stu}}$$  

(11)

$$L_{\text{dist}} = D_{\text{KL}}(\text{Softmax}(s_{\text{tea}})||\text{Softmax}(s_{\text{tea}}))$$  

(12)

$$D_{\text{KL}}(P||Q) = \sum_{i} p_i \log \left( \frac{p_i}{Q_i} \right)$$  

(13)

The potential mistakes learned by the teacher may negatively affect the student. As a result, we also train the distilled student model to produce the correct ground truth labels in addition to the soft supervision from the teacher. Concretely, we calculate the prediction loss $L_{\text{pred}}$ between the student’s output and the ground truth labels as the hard supervision. And the loss of the student model ($L_{\text{stu}}$) is precisely the sum of soft and hard supervision loss,

$$L_{\text{stu}} = L_{\text{pred}} + L_{\text{dist}}.$$  

(14)

Finally, we transfer GRUs from the student model to the target model, and fine-tune the target model with the target dataset ($R_{\text{tar}}$) using loss term $L_{\text{tar}} = L_{\text{pred}}$. The specific process of our algorithm DistCare is presented in Algorithm 1.

**Algorithm 1 DistCare ($R_{\text{src}}, R_{\text{tar}}$)**

1. **Stage 1:** Randomly initializing parameters in Teacher Model $\text{DistCare}_{\text{tea}}$
   2. while not convergence do:
      3. Compute $y_{T,\text{src}}, s_{\text{tea}} = \text{DistCare}_{\text{tea}}(R_{\text{src}})$
      4. Compute $L_{\text{tea}} = \text{MSE}(y_{T,\text{src}}, y_{T,\text{src}})$
      5. Update parameters of DistCare$_{\text{tea}}$ by optimizing $L_{\text{tea}}$ using back-propagation
   end while
3. **Stage 2:** Randomly initializing parameters in Student Model $\text{DistCare}_{\text{stu}}$
   4. while not convergence do:
      5. Compute $y_{T,\text{src}}, s_{\text{tea}} = \text{DistCare}_{\text{stu}}(R_{\text{src}})$
      6. Compute $L_{\text{pred}} = \text{MSE}(y_{T,\text{src}}, y_{T,\text{src}})$
      7. Compute $L_{\text{dist}} = D_{\text{KL}}(\text{Softmax}(s_{\text{tea}})||\text{Softmax}(s_{\text{tea}}))$
      8. Compute $L_{\text{stu}} = L_{\text{pred}} + L_{\text{dist}}$
      9. Update parameters of DistCare$_{\text{stu}}$ by optimizing $L_{\text{stu}}$ using back-propagation
   end while

15. **Stage 3:** Transfer parameters of shared GRUs from DistCare$_{\text{stu}}$ to Target Model DistCare$_{\text{tar}}$, and randomly initializing other parameters in DistCare$_{\text{tar}}$
   16. while not convergence do:
      17. Compute $y_{T,\text{tar}} = \text{DistCare}_{\text{stu}}(R_{\text{tar}})$
      18. Compute $L_{\text{tar}} = \text{MSE}(y_{T,\text{tar}}, y_{T,\text{tar}})$
      19. Update parameters of DistCare$_{\text{tar}}$ by optimizing $L_{\text{tar}}$ using back-propagation
   end while

5 EXPERIMENT

We conduct the experiments by leveraging publicly available online PhysioNet Source Dataset [37] to enhance the LOS (Length of Stay) prediction on COVID-19 datasets [18, 44]. To further verify the scalability of DistCare when performing different clinical prediction tasks on different EMR datasets, we also conduct an additional mortality prediction experiment on the end-stage renal disease (ESRD) dataset. Our code and the visualization system are available at https://github.com/Accountable-Machine-Intelligence/DistCare.
5.1 Data Description

5.1.1 COVID-19 Target Dataset from Tongji Hospital (TJH), China. We take the COVID-19 dataset [44] as the target dataset and perform the LOS prediction. The medical information of all patients collected between 10 January and 18 February 2020 was used for model training. The average age of the patients was 58.8 years old, and 59.7% were male. Of the 375 cases included in the subsequent analysis, 201 recovered from COVID-19 and was discharged from the hospital, while 174 unfortunately died. Statistics of the source dataset and target dataset are listed in Table 2. Statistics of the LOS are listed in Table 3. The distribution of days to the outcome for records is shown in Figure 2. Medical features recorded in TJH target dataset are listed in Table 4.

5.1.2 COVID-19 Target Dataset from HM Hospitals (HMH), Spain. HMH [18] is released by HM Hospitals containing 2,310 anonymous patients diagnosed with COVID-19 or to be confirmed. These data collect various interactions during the treatment of COVID-19, including detailed information about the diagnosis, treatment, admission, steps through the ICU, and outcomes, discharge or death. We selected the patients who have at least one record of lab tests. After screening, there are 1,891 patients, and 303 patients died. The distribution of length of stay in HM Hospital is shown in Figure 3. Medical features recorded in HMH target dataset are listed in Table 5.

5.1.3 PhysioNet Source Dataset. We take the publicly available PhysioNet Dataset [37] as the source dataset and pre-train the medical feature embedding based on the Sepsis prediction. This dataset is sourced from ICU patients in two separate U.S. hospital systems. These data were collected over the past decade with approval from the appropriate Institutional Review Boards, and are labeled by Sepsis-3 clinical criteria. The cleaned dataset consists of 40,336 patients and consists of a combination of hourly vital sign summaries (e.g., heart rate, systolic blood pressure), laboratory values (e.g., chloride, glucose). In particular, the data contained 34 clinical variables: 8 vital sign variables and 26 laboratory variables. The statistics of the datasets are presented in Table 2. Medical features recorded in the PhysioNet source dataset are listed in Table 4.

5.1.4 Additional Experiment: End-Stage Renal Disease Target Dataset. We take the ESRD dataset as the extended target dataset and perform the mortality prediction. Nowadays, many people suffer from
End-Stage Renal Disease (ESRD) in the world [21, 41]. They face severe life threats and need lifelong treatments with periodic visits to the hospitals for various tests (e.g., blood routine examination). The whole procedure needs a dynamic patient health risk prediction to help patients recover smoothly and prevent adverse outcomes, based on the medical records collected along with the visits. This task is defined as a binary classification task of predicting a patient’s death in one year.

In this study, all ESRD patients who received therapy from January 1, 2006, to March 1, 2018, in a real-world hospital are included to form this dataset. There are 1196 records with positive labels (i.e., died within 12 months) and 10,804 records with negative labels. The core task is to learn the patient’s health status representation and perform the mortality prediction at each record. We drop the patients whose all entries of one feature are missing and select the observed features in more than 60% of patients’ records. For missing values, we fill the missing front cells with the data backward to prevent future information leakage. If the patient’s backward record is missing, we impute it with the patient’s first front observed record. The cleaned dataset consists of 656 patients and 13,091 visits. The statistics of the ESRD dataset are presented in Table 2. Medical features recorded in the ESRD target dataset are listed in Table 6.

### 5.2 Experimental Setup

#### 5.2.1 Evaluation Preparation

Due to the limited amount of data, 10-fold Cross-Validation is employed on the prediction task. The numbers in parentheses (Table 7) denote the standard deviation of 10-fold cross-validation. We assess the performance of the regression task (i.e., LOS Prediction) using Mean Square Error (MSE) and Mean Absolute Error (MAE). Specifically,

\[
MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \quad (15)
\]

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|, \quad (16)
\]

Alternatively, for the binary classification task (i.e., mortality prediction), we assess performance using the Area Under Receiver Operating Characteristic Curve (AUROC), Area Under Precision-Recall Curve (AUPRC), and the Minimum of Precision and Sensitivity Min\(\bar{Se}, \bar{P}\).

#### 5.2.2 Baseline Approaches

We introduce several deep-learning-based models as our baseline approaches without additional labeled data or external ontology resources.

- GRU [7] is the basic Gated Recurrent Unit network.
- StageNet (WWW’20) [13] extracts disease stage information from patient data and integrate it into risk prediction.
To train the model, we use Adam [23] with the batch size of 256, and the learning rate is set to $1 \times 10^{-3}$. To fairly compare different approaches, the hyper-parameters of the baseline models are fine-tuned by the grid-searching strategy.

### Table 8: Additional Experiment: Mortality Prediction Performance on ESRD Dataset.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AUPRC</th>
<th>AUROC</th>
<th>Min(Se, P+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRU</td>
<td>0.7142 (0.0883)</td>
<td>0.8094 (0.0547)</td>
<td>0.6668 (0.0544)</td>
</tr>
<tr>
<td>StageNet</td>
<td>0.7205 (0.0657)</td>
<td>0.8240 (0.0337)</td>
<td>0.6911 (0.0364)</td>
</tr>
<tr>
<td>ConCare</td>
<td>0.7291 (0.0827)</td>
<td>0.8259 (0.0456)</td>
<td>0.6784 (0.0573)</td>
</tr>
<tr>
<td>T-LSTM</td>
<td>0.7120 (0.0841)</td>
<td>0.8066 (0.0628)</td>
<td>0.6702 (0.0512)</td>
</tr>
<tr>
<td>TimeNet</td>
<td>0.6328 (0.0310)</td>
<td>0.7311 (0.0262)</td>
<td>0.5926 (0.0194)</td>
</tr>
<tr>
<td>AMT</td>
<td>0.5759 (0.0933)</td>
<td>0.6940 (0.0741)</td>
<td>0.5661 (0.0516)</td>
</tr>
<tr>
<td>AttBiGRU</td>
<td>0.6573 (0.0776)</td>
<td>0.7514 (0.0589)</td>
<td>0.6306 (0.0644)</td>
</tr>
<tr>
<td>DistCare$_{stu}$</td>
<td>0.7414 (0.0692)</td>
<td>0.8263 (0.0427)</td>
<td>0.6723 (0.0523)</td>
</tr>
<tr>
<td>DistCare</td>
<td>0.7614 (0.0584)</td>
<td>0.8361 (0.0385)</td>
<td>0.7046 (0.0353)</td>
</tr>
</tbody>
</table>

- ConCare (AAAI’20) [29] embeds the feature sequences separately and uses the self-attention to model dynamic features and static baseline information.
- T-LSTM (SIGKDD’17) [1] handles irregular time intervals by time decay mechanism. We modify it into a supervised learning model.
- AMT (WWW’20) [46] combines multi-task learning and transfer learning framework to allow knowledge to be shared across domains and tasks. ²
- AttBiGRU (BIBM’19) [39] proposes a general transfer learning strategy which can enable models to make clinical prediction acrossing diverse EHRs datasets. ²
- TimeNet (IJCAI-Workshop’18) [15] maps variable-length clinical time series to fixed-dimensional feature vectors separately, and acts as an off-the-shelf feature extractor. ³
- DistCare$_{stu}$ is the proposed DistCare without distillation from the teacher model.

#### 5.2.3 Experiment Environment

The experiment is conducted on a machine equipped with CPU: Intel Xeon E5-2630, 256GB RAM, and GPU: NVIDIA Titan X. The code is implemented based on Pytorch 1.5.0. To train the model, we use Adam [23] with the batch size of 256, and the learning rate is set to $1 \times 10^{-3}$. To fairly compare different approaches, the hyper-parameters of the baseline models are fine-tuned by the grid-searching strategy.

---

²Transfer-learning-based baseline models are pre-trained on the PhysioNet ICU source dataset [37] https://physionet.org.

---

Figure 4: Prediction performance on COVID-19 Tongji Hospital dataset under different training data volume.

### 5.3 Experiment Results

As is shown in Table 7, DistCare consistently outperforms both transfer-based and non-transfer-based baselines, demonstrating the ability of DistCare to learn a robust representation. Concretely, for the COVID-19 LOS prediction task, compared to the best state-of-the-art model ConCare, DistCare achieves 6% lower MSE, 5% lower MAE relatively on COVID-19 TJH dataset, and achieves 9.6% lower MSE, 6.4% MAE relatively on HMH dataset. Compared to StageNet, another best baseline method on MAE, DistCare achieves 27% lower MSE relatively on TJH dataset, and 14.8% lower MSE relatively on HMH dataset. For the extended ESRD mortality prediction task, compared to ConCare, DistCare also achieves a 4.4% higher AUPRC, a 1.24% higher AUROC, and a 3.86% higher min(Se, P+) relatively.

- **Effectiveness of Transfer Learning:** By comparing the models with and without transfer mechanism (i.e., ConCare and DistCare), we can conclude that utilizing knowledge from existing publicly available EMR can significantly promote the prediction performance of models on both tasks, indicating the effectiveness of the feature-specific transfer learning mechanism. Moreover, DistCare also shows a better performance than other transfer-learning-based methods. Though these models employ the transferring mechanism, our model DistCare executes a more adaptive and reasonable feature-specific transfer.
5.4 Observation on COVID-19 TJH: Varying the Size of Training Set

We evaluate whether DistCare can reach a robust performance even under insufficient data volume on the COVID-19 dataset. Several training datasets with different amounts of data (i.e., 90/80/50/20/10% of the whole dataset) are created by adjusting cross-validation experiments. Figure 4 shows the mean square error (MSE) for LOS prediction on validation sets under different training data amount.

The MSEs of all models rise as the training data volume decreases, and DistCare consistently outperforms all baselines on all training sets with different sizes. The performance gap between DistCare and the baselines is more considerable in smaller datasets, which indicates the capability of distilled transfer mechanism to alleviate the data insufficiency problem. When we adopt only 10% of the whole dataset as the training set, which is the smallest of all experiment settings, DistCare demonstrated significantly better performance than the best baseline ConCare. Specifically, DistCare achieves an MSE of 336.3897, while the baseline models ConCare and AMT achieve 491.2162 and 446.7076, showing 31.5% and 24.7% relative improvement, respectively.

5.5 Observation on COVID-19 TJH: Patient Health Representation Learning

In this subsection, we make observations on the COVID-19 dataset to evaluate whether DistCare can extract a robust representation of patient health status. We divide the regression labels \( y \) into four classes according to the severity of patients’ health status:

- **Very Low**: \( y < 7 \), Discharge in 7 Days.
- **Low**: \( 7 \leq y < 35 \), Discharge over 7 Days.
- **High**: \( 35 \leq y < 63 \), Death over 7 Days.
- **Critical**: \( y \geq 63 \), Death in 7 Days.

We plot the confusion matrix of GRU and DistCare respectively, presented in Fig. 5. GRU seems unable to distinguish the records with extremely high risk from other patients with death outcomes. It can only vaguely inform doctors whether the patient is in a dangerous state, but can not carry out different levels of warnings.

In comparison, DistCare can correctly predict more Critical patients, demonstrating a better ability to distinguish the different severity levels of patients’ records. This makes it possible to conduct personalized diagnosis and treatment among different patients and monitor every patient’s health condition dynamically.

To make further observations, we visualize patients’ health status embeddings obtained from DistCare in Figure 6. Each colored dot represents a patient’s...
which is regarded as a strong indication of deterioration, leading to the rapid rise of predicted health risk. For the COVID-19 pandemic, rapid and effective triage is critical for early treatment and effective hospital resource allocation. Through the LOS prediction, doctors can perform a more accurate assessment of the patient’s future outcomes, giving more individualized treatments for patients. This ensures that the patients can receive targeted early treatment and remedies on deteriorating biomarkers. The anonymous case-patient is a 64-year-old male. As shown in Fig. 7, at the beginning of the treatment (1st-6th records), the patient’s health status appears to be improving. However, from day-4 to day-8 (9th-25th records), the patient’s condition is predicted to be deteriorating rapidly. The patient’s predicted health risk leaps from 30 (5th record) to 60 (25th record). The ground truth label shows that the patient dies unfortunately in 15 days. During this period, Lymphocyte Count, Total Protein and Platelet Large Cell Ratio (P-LCR) are strongly focused on by DistCare, which are plotted in Fig. 7. All these biomarkers rise or decrease acutely during this period, which is regarded as a strong indication of deterioration, leading to the rapid rise of predicted health risk. For the COVID-19 pandemic, rapid and effective triage is critical for early treatment and effective hospital resource allocation. Through the LOS prediction, doctors can perform a more accurate assessment of the patient’s future outcomes, giving more individualized treatments for patients. This ensures that the patients can receive targeted early treatment and remedies on deteriorating biomarkers.

6 CONCLUSION
In this paper, we propose a distilled transfer learning framework, DistCare, to perform the length of stay prediction for patients with COVID-19. In order to embed the medical features robustly, the model is trained to imitate the teacher model’s medical embedding behavior via soft distillation supervision. The experimental results on real-world COVID-19 datasets show that DistCare consistently outperforms several competitive baseline methods, and may benefit the intelligent prognosis for tackling future emerging infectious diseases.

ACKNOWLEDGMENTS
This work is supported by the National Natural Science Foundation of China (61772045), the Project 2019BD005 PKU-Baidu fund, and Peking University Medicine Seed Fund for Interdisciplinary Research (BMU2020M101). WR is supported by ORCA PRF Project (EP/R026173/1).

REFERENCES

Figure 7: Case study: The rising LOS health risk score curve of an anonymous case-patient. The patient’s health status deteriorates from day-4 to day-8 (6th-20th records). The case study is available at https://github.com/Accountable-Machine-Intelligence/DistCare.


