Article

Patterns

An adaptive federated learning framework for clinical risk prediction with electronic health records from multiple hospitals

Highlights

- We propose a FL framework to address data distribution drift
- The framework separates input features by relationships to clinical outcomes
- The framework can provide clinically interpretable results

Authors

Weishen Pan, Zhenxing Xu, Suraj Rajendran, Fei Wang

Correspondence

few2001@med.cornell.edu

In brief

Data distribution drift is a key challenge when building predictive models with data from multiple institutions under the FL framework. This work proposes an adaptive FL framework to address this challenge by separating input features based on their relationships to clinical outcomes. On the tasks of predicting the onset risk of sepsis and acute kidney injury for intensive care unit patients, this framework outperforms existing FL models. It can also provide reasonable feature interpretations.

Pan et al., 2024, Patterns *5*, 100898
January 12, 2024
https://doi.org/10.1016/j.patter.2023.100898

CellPress

Patterns

CellPress
OPEN ACCESS

Article

An adaptive federated learning framework for clinical risk prediction with electronic health records from multiple hospitals

Weishen Pan,[1,2] Zhenxing Xu,[1,2] Suraj Rajendran,[3] and Fei Wang[1,2,4,*]
[1]Department of Population Health Sciences, Weill Cornell Medical College, Cornell University, New York, NY 10065, USA
[2]Institute of Artificial Intelligence for Digital Health, Weill Cornell Medical College, Cornell University, New York, NY 10065, USA
[3]Tri-Institutional Computational Biology & Medicine Program, Weill Cornell Medical College, Cornell University, New York, NY 10065, USA
[4]Lead contact
*Correspondence: few2001@med.cornell.edu
https://doi.org/10.1016/j.patter.2023.100898

THE BIGGER PICTURE With the wide use of machine learning to train clinical risk prediction with EHR data, combining data from multiple institutions can benefit model training. It is, however, usually infeasible to transfer data between institutions because of data privacy regulations. While FL is proposed to enable privacy-preserving collaboration between institutions to train predictive models, data distribution drift across different institutions makes learning a universally good "global" model very challenging. We propose an adaptive FL framework to address this challenge and evaluate it on a large-scale intensive care unit dataset to predict the onset risk of sepsis and AKI. The experiment results show that our framework outperforms FL baselines and it is also clinically interpretable.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Clinical risk prediction with electronic health records (EHR) using machine learning has attracted lots of attentions in recent years, where one of the key challenges is how to protect data privacy. Federated learning (FL) provides a promising framework for building predictive models by leveraging the data from multiple institutions without sharing them. However, data distribution drift across different institutions greatly impacts the performance of FL. In this paper, an adaptive FL framework was proposed to address this challenge. Our framework separated the input features into stable, domain-specific, and conditional-irrelevant parts according to their relationships to clinical outcomes. We evaluate this framework on the tasks of predicting the onset risk of sepsis and acute kidney injury (AKI) for patients in the intensive care unit (ICU) from multiple clinical institutions. The results showed that our framework can achieve better prediction performance compared with existing FL baselines and provide reasonable feature interpretations.

INTRODUCTION

In recent years, due to the better availability of healthcare data such as electronic health records (EHRs) and rapid advancement in artificial intelligence techniques, more and more effort has been made to mine data-driven insights for improving the quality of care delivery. Among these efforts, machine learning (ML)-based clinical risk prediction, which aims at building ML models for predicting clinical outcomes (e.g., mortality or disease onset) using observational data (e.g., EHR), has been one of the most important research topics.[1] Most of the existing studies used data from a single institution to build the predictive model, which was challenging to generalize well to other institutions with different patient demographics.[2] Aggregating data from multiple institutions can increase the training data sample size for building the model, diversify the patient population, and benefit model generalizability.[3] However, the sensitive information contained in patient data prohibits them to be shared straightforwardly due to privacy concerns.[4]

Federated learning (FL),[5] which constructs ML models collaboratively by leveraging the data from multiple local sites but without sharing them out, holds great promise in medical

applications because of their privacy-preservation design.[6,7] Classical FL updates the model parameters iteratively. At each iteration, there are two main steps: updating model parameters locally with site-specific data and transmitting these local model parameter updates to a central server for aggregation to a new set of model parameters. This strategy aims to learn a global model that can work better than the locally trained models. However, the data distributions at different local sites are usually different due to the distinct population characteristics, which makes it challenging to learn a universally good model without special considerations.[8] In a recent paper,[9] we investigated this issue and demonstrated the heterogeneous performance of FL models across different local sites for clinical risk prediction tasks. There have been existing studies (such as model-agnostic meta-learning and federated multitask learning[10–13]) trying to address this issue, but these methods are not designed specifically for medicine and are difficult to explain.

To fill this research gap, we propose an adaptive FL framework for predictive modeling of clinical risks with EHR data from multiple clinical institutions. In particular, we treat each institution as a specific domain and propose to separate the input patient features into three parts: stable, domain-specific, and conditional-irrelevant. Stable/domain-specific features are predictive of the clinical outcome, but the relationships between stable features and clinical outcome are the same across all domains, while the relationships between domain-specific features and clinical outcome are different with respect to different domains. Conditional-irrelevant features are the residual features excluding stable and domain-specific features. To account for the heterogeneity of sample distributions across different sites, we learn a specific model for each site, and these site-specific models are jointly learned. The model parameters for stable features are shared across sites and the model parameters for domain-specific features are different at different sites. Similar to other FL approaches, our framework does not share any data outside the sites they reside in during the model training process. We validated the effectiveness of our method on a large-scale real-world patient EHR corpus collected from the intensive care units (ICUs) of hundreds of hospitals, where we focused on predicting the onset risk of two critical conditions in the ICU. We demonstrated that our approach could perform better than other baseline FL approaches. The identified shared and domain-specific features are clinically interpretable.

RESULTS

Cohort characteristics

Patients' EHR used in our experiments were extracted from the eICU Collaborative Research Database (eICU-CRD)[14], which is a deidentified and publicly available dataset comprising information of patients admitted to critical care units between 2014 and 2015 across the United States, including demographics, vital sign measurements, diagnoses, and treatments. We identified the sepsis patients and the patients with acute kidney injury (AKI) based on the Sepsis-3 clinical criteria[15] and Kidney Disease Improving Global Outcomes (KDIGO),[16]respectively. Basic cohort characteristics of these patients across different hospitals are summarized in Table 1. To build models for predicting the risk of sepsis and AKI, we leveraged four types of patient information as input features including vital signs, laboratory measurements, medications, and demographics, which result in a total of 358 feature variables. For sepsis, we predicted the onset risk of sepsis in the next 6 h based on their historical data.[17] For AKI, we used data during the first 24 h from ICU admission to predict risk of AKI onset within the next 24 h.[18] The overview of this framework is illustrated in Figure 1. The details of eICU, definition of KDIGO, Sepsis-3 criteria, construction of feature vector were introduced in the Experimental procedures.

Model performance

Our method was compared with (1) a pooled model (Pooled): a global model shared across all sites trained with their combined data; (2) an individual model (Indiv): individual models, each of which is trained and tested with data from each individual site; (3) models for multi-domain learning: Indiv-L2,[19] regularized multi-task learning,[20] multi-task adversarial network,[21] and adapt to adaptation for federated learning.[22]

Individual area under the receiver operating characteristic (AUROC) calculated at each hospital (indexed from 1 to 7) are reported in Table 2, which shows that our method obtained the best results in three of seven sites for AKI prediction, and four of seven for sepsis prediction. Table 3 summarizes the micro and macro AUROC over all seven sites of different algorithms, which demonstrates that our method performs the best on both tasks. In addition, an ablation study was performed to investigate the performance of using different types of input features on building these risk prediction models, and the results are demonstrated in Table 4, which shows that laboratory findings and vital signs along with medications are more predictive than demographics for both tasks.

Interpretation

To obtain an intuitive understanding of the sample distributions across different hospitals, we visualize the patient vectors using the uniform manifold approximation and projection (UMAP) technique[23] as in Figure 2, where the patients from hospitals are colored differently. The left column of Figure 2 is the embeddings of patient vectors composed of all features, which clearly demonstrates the distribution heterogeneity (e.g., samples from hospital 5 in the AKI task are separated from other samples). The middle column of Figure 2 is the UMAP embeddings of sample vectors formed by the learned stable features, where the samples from different hospitals are really blended with each other. The right column of Figure 2 is the UMAP embeddings of the sample vectors formed by domain-specific features, from which we can observe more scattered point clouds that are specific to individual sites.

The top 10 stable and domain-specific features for AKI and sepsis prediction are reported in Table 5. The quantitative contributions of representative features to the predictions calculated by Shapely Additive exPlanations (SHAP)[24] are shown in Figures 3 and 4, where, for all subfigures, the horizontal axes represent the feature value (0 or 1 for binary features, Z score normalized value for continuous features) and the vertical axes are the Shapley values. In both figures, the top row

Table 1. Summary statistics of the demographic and outcome variables in all hospitals (indexed from 1 to 7) for AKI and sepsis prediction

	Hospital ID	1	2	3	4	5	6	7
	No. (%)	5,545	2,848	2,990	3,665	3,302	2,957	2,939
AKI	Positive	306 (5.5%)	16 0(5.6%)	286 (9.6%)	177 (4.8%)	195 (5.9%)	226 (7.6%)	251 (8.5%)
	Age							
	18–39	505 (9.1%)	355 (12.5%)	243 (8.1%)	407 (11.1%)	429 (13.0%)	313 (10.6%)	633 (21.5%)
	40–59	1,699 (30.6%)	773 (27.1%)	888 (29.7%)	1,133 (30.9%)	983 (29.8%)	928 (31.4%)	1,123 (38.2%)
	≥ 60	3,334 (60.1%)	1,715 (60.2%)	1,854 (62.0%)	2,119 (57.8%)	1,888 (57.2%)	1,713 (57.9%)	1,155 (39.3%)
	Sex							
	Female	2,452 (44.2%)	1,232 (43.3%)	1,316 (44.0%)	1,741 (47.5%)	1,429 (43.3%)	1,198 (40.5%)	1,251 (42.6%)
	Male	3,093 (55.8%)	1,616 (56.7%)	1,673 (56.0%)	1,924 (52.5%)	1,873 (56.7%)	1,756 (59.4%)	1,684 (57.3%)
	Ethnicity							
	Caucasian	4,151 (74.9%)	2,761 (96.9%)	1,953 (65.3%)	3,192 (87.1%)	3,072 (93.0%)	2,542 (86.0%)	1,301 (44.3%)
	African American	813 (14.6%)	43 (1.5%)	909 (30.4%)	272 (7.4%)	47 (1.4%)	131 (4.4%)	1,513 (51.5%)
	Hispanic	397 (7.2%)	26 (0.9%)	0 (0.0%)	11 (0.3%)	44 (1.3%)	2 (0.1%)	32 (1.1%)
	Asian	72 (1.3%)	2 (0.1%)	28 (0.9%)	31 (0.8%)	11 (0.3%)	46 (1.6%)	13 (0.4%)
	Others	112 (2.0%)	16 (0.6%)	100 (3.3%)	159 (4.3%)	128 (3.9%)	236 (8.0%)	80 (2.7%)
Sepsis	No. (%)	5,919	2,996	3,212	2,748	3,578	2,276	3,344
	Positive	89 (1.5%)	135 (4.5%)	187 (5.8%)	123 (4.5%)	21 (0.1%)	459 (20%)	13 (0.4%)
	Age							
	18–39	523 (8.8%)	359 (12.0%)	259 (8.1%)	297 (10.8%)	443 (12.4%)	235 (10.3%)	694 (20.8%)
	40–59	1,813 (30.6%)	835 (27.9%)	947 (29.5%)	899 (32.7%)	1,057 (29.5%)	694 (30.5%)	1,305 (39.0%)
	≥60	3,576 (60.4%)	1,797 (60.0%)	2,000 (62.3%)	1,549 (56.4%)	2,076 (58.0%)	1,344 (59.1%)	1,316 (39.4%)
	Sex							
	Female	2,629 (44.4%)	1,293 (43.2%)	1,404 (43.7%)	1,234 (44.9%)	1,541 (43.1%)	936 (41.1%)	1,424 (42.6%)
	Male	3,290 (55.6%)	1,702 (56.8%)	1,807 (56.3%)	1,514 (55.1%)	2,037 (56.9%)	1,338 (58.8%)	1,916 (57.3%)
	Ethnicity							
	Caucasian	4,394 (74.3%)	2,896 (96.7%)	2,055 (64.0%)	1,757 (63.9%)	3,329 (93.0%)	1,942 (85.6%)	1,439 (43.0%)
	African American	899 (15.2%)	44 (1.5%)	1,019 (31.7%)	805 (29.3%)	52 (1.5%)	100 (4.4%)	1,766 (52.8%)
	Hispanic	428 (7.2%)	36 (1.2%)	0 (0.0%)	0 (0.0%)	48 (1.3%)	2 (0.1%)	38 (1.1%)
	Asian	82 (1.4%)	3 (0.1%)	31 (1.0%)	38 (1.4%)	11 (0.3%)	43 (1.9%)	13 (0.4%)
	Others	116 (2.0%)	17 (0.6%)	107 (3.3%)	148 (5.4%)	138 (3.9%)	189 (8.3%)	88 (2.6%)

corresponds to the plots for stable features, and the bottom row are the plots for domain-specific features. These figures show similar curves for stable features across different hospitals, suggesting that the relationships between these features and predicted outcomes are similar across hospitals. In contrast, the curves for domain-specific features are much more heterogeneous (e.g., features such as ASPIRIN play a fairly important role for AKI prediction at hospital 3, but not in others, and the contributions of LISPRO are positive for AKI prediction at hospital 6, but negative for other hospitals, suggesting that these domain-specific features do have distinct effects for individual sites.

DISCUSSION

The main contribution of this paper is the development of an adaptive FL framework to handle the data distribution discrepancies across different sites in FL setting. Our framework splits the input features into stable, domain-specific, and conditional-irrelevant parts. This procedure effectively teases out the shared and specific factors contributing to the prediction of certain clinical outcomes, which can further explain the impact of distribution heterogeneity to clinical risk prediction tasks.

The effectiveness of our proposed framework was evaluated on the tasks of predicting the onset risk of sepsis and AKI in critical care setting,[25,26] where our model has demonstrated better quantitative performance over a set of state-of-the-art baselines. Such performance improvement could be coming from (1) a diverse set of information, including demographics, lab tests, vital signs, and medications, were incorporated as input features to build the ML models. This captures the patient characteristics more comprehensively compared with models only using certain types of patient information.[18,27] We also demonstrated that different types of information play different roles for clinical risk prediction in Table 4. (2) The proposed framework learns a set of site-specific

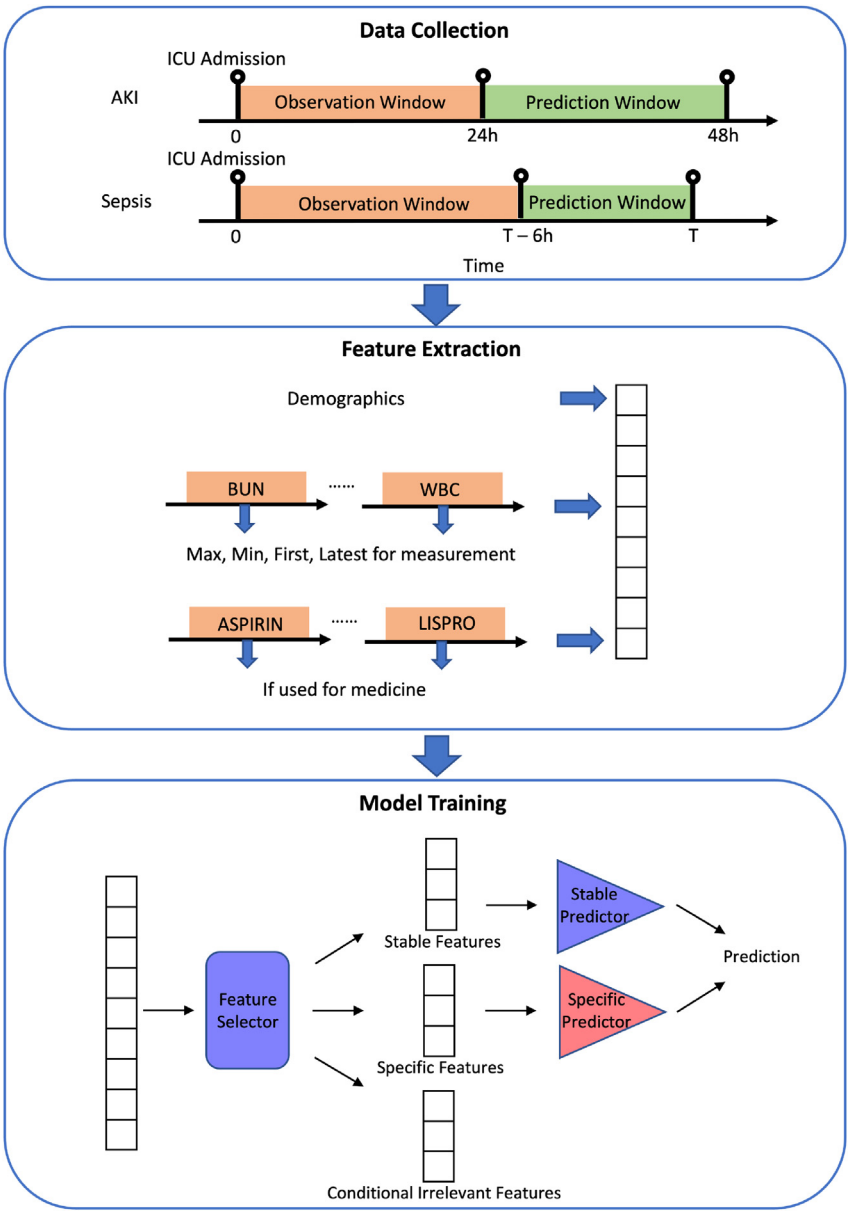

Figure 1. The overview of our proposed framework

In the data collection process, the positive and negative samples are identified based on the clinical criteria. Subsequently, the prediction window and data observation window are selected. In the data pre-processing process, a feature vector will be constructed from different types of EHR data within the observational window. Then an adaptive FL-based prediction model is built on the features by separating them into stable, domain-specific and conditional-irrelevant features and treating them accordingly in the predictor.

models collaboratively in a privacy-preserving way of not exposing local data, which also effectively accounts for the distribution heterogeneity of samples across different sites.

In addition to superior quantitative performance, the identification of stable, domain-specific, and conditional-irrelevant features greatly helps model interpretability. With the SHAP technique, we showed in Figures 3 and 4 that the learned stable features contribute similarly on predicting the clinical outcomes across different hospitals, while prediction contributions from the learned domain-specific features vary greatly from hospital to hospital. The learned stable and domain-specific features shown in Table 5 also make clinical sense. In the task of AKI risk prediction, the level of creatinine is a critical indicator of the kidney function and it is used to diagnose AKI,[16] and it has been identified as an important stable feature by our model. In addition, there are several blood pressure-related stable features, and it has been suggested in prior literature that blood pressure is related to AKI etiology.[28] Other stable features such as blood urea nitrogen and platelets have also been reported to be associated with AKI.[29,30] There is also a medication FUROSEMIDE in the stable feature list, which is a diuretic whose use could be associated with kidney function decline.[31] In contrast, several medications are identified as domain-specific features, which could be due to practice variance and resource availability at different hospitals. Similar observations can also be obtained from the sepsis prediction task. And the identified stable features have been previously reported as risk factors for sepsis in existing studies, such as body mass index (BMI),[32] hematocrit,[33] sodium,[34] and blood pressure.[35]

Our study has several limitations. (1) Only structured information within the EHR was leveraged in our empirical study. The unstructured portion of EHR, such as clinical notes, contains important patient information as well and can further boost the prediction performance. (2) Two particular tasks, AKI and sepsis onset prediction in critical care, were investigated to evaluate the effectiveness of our proposed model. In the future we plan to implement our framework on more clinical risk prediction tasks of different types to understand its full potential.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Fei Wang (few2001@med.cornell.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The data can be requested and downloaded at: https://physionet.org/content/eicu-crd/2.0/.[36] Our source code is available at GitHub (https://github.com/adap-fed-ehr-code/adap-fed-ehr) and has been archived at Zenodo.[37]

Table 2. AUROC performance of AKI and sepsis prediction for all hospitals (indexed from 1 to 7)

	Hospital ID	1	2	3	4	5	6	7
	Pooled	0.719 (0.027)	0.722 (0.019)	0.801 (0.026)	0.723 (0.045)	0.674 (0.039)	0.761 (0.022)	0.775 (0.031)
	Indiv	0.629 (0.034)	0.692 (0.042)	0.757 (0.028)	0.669 (0.052)	0.645 (0.045)	0.711 (0.024)	0.729 (0.040)
	Indiv-L2	0.693 (0.030)	0.700 (0.057)	0.762 (0.028)	0.685 (0.038)	0.651 (0.059)	0.727 (0.024)	0.747 (0.022)
AKI	RMTL	0.736 (0.022)	0.733 (0.043)*	0.805 (0.028)	0.727 (0.056)	0.677 (0.051)	0.736 (0.036)	0.799 (0.042)*
	MAN	0.624 (0.032)	0.678 (0.028)	0.743 (0.026)	0.645 (0.052)	0.615 (0.046)	0.697 (0.048)	0.720 (0.036)
	APPLE	0.736 (0.033)	0.718 (0.043)	0.821 (0.027)*	0.710 (0.043)	0.660 (0.061)	0.790 (0.051)*	0.782 (0.040)
	Ours	0.754 (0.025)*	0.728 (0.049)	0.818 (0.034)	0.735 (0.030)*	0.679 (0.026)*	0.777 (0.037)	0.787 (0.051)
Sepsis	Pooled	0.833 (0.028)	0.687 (0.022)	0.778 (0.027)	0.777 (0.024)	0.614 (0.041)	0.761 (0.027)	0.810 (0.027)
	Indiv	0.736 (0.048)	0.696 (0.025)	0.772 (0.034)	0.776 (0.042)	0.637 (0.033)	0.761 (0.028)	0.784 (0.056)
	Indiv-L2	0.805 (0.047)	0.708 (0.038)	0.790 (0.045)	0.777 (0.025)	0.633 (0.042)	0.772 (0.026)	0.830 (0.023)
	RMTL	0.798 (0.061)	0.704 (0.036)	0.761 (0.049)	0.771 (0.031)	0.626 (0.032)	0.736 (0.023)	0.812 (0.044)
	MAN	0.745 (0.063)	0.709 (0.031)*	0.771 (0.056)	0.752 (0.027)	0.609 (0.033)	0.761 (0.012)	0.801 (0.037)
	APPLE	0.802 (0.034)	0.675 (0.052)	0.772 (0.027)	0.807 (0.029)*	0.627 (0.067)	0.778 (0.019)	0.853 (0.061)*
	Ours	0.861 (0.049)*	0.701 (0.034)	0.825 (0.041)*	0.803 (0.028)	0.659 (0.046)*	0.790 (0.025)*	0.841 (0.057)

APPLE, adaptation for federated learning; MAN, multi-task adversarial network; RMTL, regularized multi-task learning. Asterisks denote the best-performing approaches within the respective columns.

Study cohort

The EHRs employed in this study are from the eICU-CRD,[14] which is a deidentified and publicly available dataset that meets the safe harbor provision of the U.S. Health Insurance Portability and Accountability Act. The eICU-CRD is a multi-center database sourced from the Philips eICU program, a telemedicine initiative where healthcare workers remotely monitor acutely ill patients. It comprises 200,859 patient unit encounters for 139,367 unique patients admitted between 2014 and 2015 across the United States. The patient information includes demographics, vital sign measurements, care plan documentation, severity of illness measures, diagnosis information, treatment information, and more.

Ethics statement

The eICU database was accessed via the PhysioNet platform. Access to the database was approved after completing the Collaborative Institutional Training Initiative program "Data or Specimens Only Research" (certificate ID: 33510902), as well as signing the data usage agreement of the PhysioNet Review Board. The study was exempt from approval from the institutional review board of the Massachusetts Institute of Technology because of the retrospective design, lack of direct patient intervention, and the security schema, for which the re-identification risk was certified as meeting safe harbor standards by an independent privacy expert (Privacert) (Health Insurance Portability and Accountability Act Certification no. 1031219-2). The institutional review board of the Massachusetts Institute of Technology waived the need for informed consent for the same reason.

Table 3. Macro/micro AUROC performance of AKI and sepsis prediction

	AKI		Sepsis	
	macro	micro	macro	micro
Pooled	0.740 (0.023)	0.738 (0.019)	0.752 (0.030)	0.805 (0.024)
Indiv	0.690 (0.021)	0.680 (0.022)	0.737 (0.043)	0.775 (0.021)
Indiv-L2	0.709 (0.017)	0.706 (0.018)	0.759 (0.034)	0.808 (0.018)
RMTL	0.744 (0.023)	0.742 (0.021)	0.744 (0.024)	0.811 (0.023)
MAN	0.674 (0.024)	0.673 (0.019)	0.736 (0.037)	0.784 (0.028)
APPLE	0.746 (0.018)	0.748 (0.018)	0.759 (0.023)	0.824 (0.012)
Ours	0.754 (0.019)*	0.753 (0.017)*	0.783 (0.025)*	0.826 (0.016)*

APPLE, adaptation for federated learning; MAN, multi-task adversarial network; RMTL, regularized multi-task learning. Asterisks denote the best-performing approaches within the respective columns.

The study was conducted following the Declaration of Helsinki. All methods used in this study were performed in accordance with the relevant guidelines and regulations.

Data preparation and preprocessing

In this study, each ICU stay is considered as an individual data sample. In cases where a patient has multiple ICU stays, we only consider the first ICU stay to prevent any potential information leakage. We focus on predicting the onset of AKI and sepsis. For both AKI and sepsis, the task is to predict the risk of disease onset during the prediction window using data collected during the data observation window, as illustrated in Figure 1. To ensure that our settings align with existing clinical research,[8,38,39] we tailor our prediction and data observation windows differently for AKI and sepsis.

For AKI prediction, we used data from the first 24 h from ICU admission to predict disease risk within the next 24 h. AKI is defined by KDIGO.[16] It is defined as any of the following.

(1) Increase in serum creatinine by $\geq$0.3 mg/dL ($\geq$26.5 μmol/L) within 48 h; or
(2) Increase in serum creatinine to $\geq$1.5 times baseline, which is known or presumed to have occurred within the prior 7 days; or
(3) Urine volume <0.5 mL/kg/h for 6 h.

We applied the definition above to all the patients with available lab test records within 48 h after ICU admission. Positive samples are samples that are diagnosed as AKI in the prediction window while negative samples are samples that are not diagnosed as AKI. We included the patients with end-stage renal disease or those on dialysis, as we aimed to predict AKI in all situations.

For sepsis prediction, we aim to predict the onset of sepsis in the next 6 h based on their historical data after ICU admission. The onset of sepsis is determined in

Table 4. Macro AUROC performance of AKI and sepsis prediction using different kinds of medical information

	AKI	Sepsis
Overall	**0.754 (0.019)***	**0.783 (0.025)***
Demographic	0.582 (0.007)	0.606 (0.011)
Lab and vital signs	0.695 (0.015)	0.735 (0.019)
Medication	0.703 (0.013)	0.729 (0.022)

Asterisks denote the best-performing approaches within the respective columns.

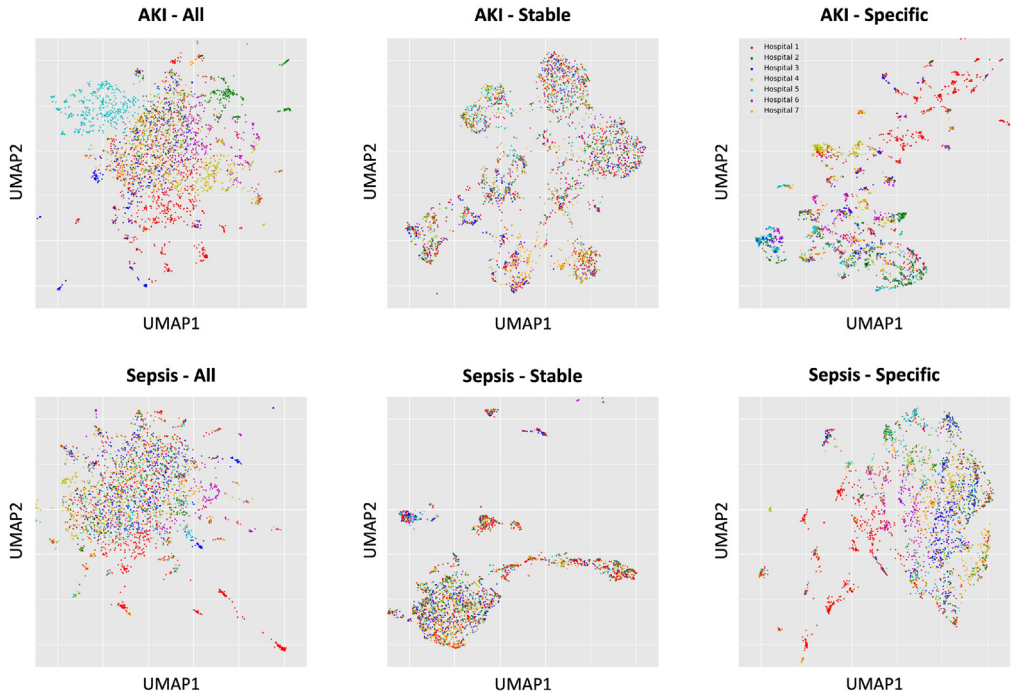

Figure 2. UMAP visualization of the features

The top row of subfigures displays UMAP visualizations for the AKI prediction task, depicting all features, stable features, and specific features from left to right. And the down row displays UMAP visualizations for the sepsis prediction task. In each subfigure, individual data points represent patients, while distinct colors are used to denote different hospitals.

accordance with the Sepsis-3 clinical criteria.[15] For each septic patient, we specified the following three time points to define the onset time t_{sepsis} of sepsis.

(1) $t_{Suspicion}$: Clinical suspicion of infection identified as the earlier timestamp of intravenous (IV) antibiotics and blood cultures within a given time interval. If IV antibiotics were given first, then the cultures must have been obtained within 24 h. If cultures were obtained first, then IV antibiotic must have been ordered within 72 h. In either case, IV antibiotics must have been administered for at least 72 consecutive hours. Note that, if there are not enough culture data, the infection can be identified according to documented diagnosis. For example, in eICU-CRD, microbiology data were not well populated due to the limited availability of microbiology interfaces; instead, infection was identified according to documented diagnosis.
(2) t_{SOFA}: Occurrence of organ failure as identified by a 2-point increase in the Sequential Organ Failure Assessment (SOFA) score within a 24-h period (SOFA score ≥ 2).
(3) t_{Sepsis}: Onset of sepsis identified as the earlier of $t_{suspicion}$ and t_{SOFA} as long as t_{SOFA} occurred no more than 24 h before or 12 h after $t_{suspicion}$.

Positive samples are samples that are diagnosed as sepsis, while controls are samples that are not diagnosed as sepsis. For positive samples, we identify the earliest time point when the aforementioned criteria are met as the onset time. For negative samples, we randomly select an index time from the distribution of onset times of the sepsis patients. Subsequently, we define the observation window as the period extending from ICU admission to 6 h prior to the onset time (for positive samples) and the index time (for negative samples).

Finally, we selected the top 7 hospitals with the most patients and available clinical records to define AKI and sepsis. The label distributions and demographics in all hospitals are shown in Table 1.

After determining the labels of the samples and the corresponding observational window, the input features are extracted from the clinical data within the observation window for predicting disease onsets. The EHR data include vital signs, laboratory tests, medications, and demographic variables. The entire feature list is in Tables S1, S2 and S3.

(1) For vital signs and laboratory tests, we extracted the earliest, latest, maximal, and minimal values within the observational window of different vital signs and laboratory tests. including heart rate, temperature, and the count of the white blood cells. For urine, only the summation is calculated. There are 29 different vital signs and laboratory tests, resulting in 113 features.
(2) For demographics, gender, age, ethnicity, and BMI were extracted in addition to a feature indicating whether the patient underwent elective surgery during the admission. There are eight demographic features.
(3) For medications, the medications were aggregated by ingredient. For each medication, whether the medication is used within the observational window was extracted as the feature. The dosage was not considered in this study. There are 237 medication features.

Adaptive prediction model

In the following section, we use capitalized/lower-case letters in italics to represent a variable/value of the variable. We use capitalized/lower-case letters in boldface to represent a variable set/values of the variable set. We represent the input feature set as $\mathbf{X} = \{X(1), \ldots, X(K)\}$ and the outcome as Y. Suppose there are M different sites, each with N_m samples (m is the index of the site). The data samples we observe are $\{\mathbf{x}_n^m, y_n^m\}, n = 1, \ldots, N_m$, $m = 1, \ldots, M$, where $\mathbf{x}_n^m \in \mathbb{R}^K$ is the input feature vector of the n -th sample at m-th site and y_n^m is its ground-truth outcome.

The overall architecture of our proposed model is shown in Figure 5. A feature separator $\mathcal{F}$ is utilized to separate $\mathbf{X}$ into three subsets: $\mathbf{S}$ representing stable features, $\mathbf{D}$ for domain-specific features, and $\mathbf{C}$ for conditional-irrelevant features. $\mathcal{F}$ comprises two cascaded stochastic gates,[40] namely $\mathcal{S}_1$ and $\mathcal{S}_2$. $\mathcal{S}_1$ initially selects $\mathbf{S}$ and $\mathbf{D}$ from $\mathbf{C}$ as a whole, and subsequently, $\mathcal{S}_2$ distinguishes between $\mathbf{S}$ and $\mathbf{D}$. It is worth noting that $\mathcal{F}$ is shared across the sites, which means the separation of $\mathbf{S}$, $\mathbf{D}$ and $\boldsymbol{C}$ is the same across the sites. After feature

Table 5. The list of top-10 stable and domain-specific features for AKI prediction

		Name	Type	Description
AKI	stable	bun_first	continuous	the first value of blood urea nitrogen (BUN) in the data observation window
		age	continuous	the age of the patient at admission
		sysbp_first	continuous	the first value of systolic blood pressure in the data observation window
		platelet_first	continuous	the first value of platelet in the data observation window
		furosemide	binary	the usage of furosemide
		creatinine_first	continuous	the first value of creatinine in the data observation window
		tempc_min	continuous	the minimal value of temperature in the data observation window
		meanbp_first	continuous	the first value of the mean of systolic blood pressure and diastolic blood pressure in the data observation window
		sysbp_max	continuous	the maximal value of systolic blood pressure in the data observation window
		resprate_last	continuous	the latest value of respirator rate in the data observation window
	specific	glucose	binary	the usage of glucose
		nitroglycerin	binary	the usage of nitroglycerin
		aspirin	binary	the usage of aspirin
		lispro	binary	the usage of lispro
		heparin	binary	the usage of heparin
		pantoprazole	binary	the usage of pantoprazole
		glucose_last	continuous	the latest value of glucose in the data observation window
		aspart	binary	the usage of aspart
		docusate	binary	the usage of docusate
		chlorhexidine	binary	the usage of chlorhexidine

(*Continued on next page*)

Table 5. *Continued*

		Name	Type	Description
Sepsis	stable	piperacillin	binary	the usage of piperacillin
		vancomycin	binary	the usage of vancomycin
		BMI	continuous	BMI
		spo2_max	continuous	the maximal value of oxygen saturation in the data observation window
		bg_paco2_first	continuous	the first value of partial pressure of carbon dioxide in the data observation window
		race_black	continuous	whether the ethnicity of the patient is African American
		hematocrit_last	continuous	the latest value of hematocrit in the data observation window
		sysbp_first	continuous	the first value of systolic blood pressure in the data observation window
		sodium_last	continuous	the latest value of sodium in the data observation window
		diasbp_min	continuous	the minimal value of diastolic blood pressure in the data observation window
	specific	chlorhexidine	binary	the usage of chlorhexidine
		glucose	binary	the usage of glucose
		glucagon	binary	the usage of glucagon
		nitroglycerin	binary	the usage of nitroglycerin
		ondansetron	binary	the usage of ondansetron
		heparin	binary	the usage of heparin
		bands_last		the latest value of bands in the data observation window
		aspirin	binary	the usage of aspirin
		lispro	binary	the usage of lispro
		fentanyl	binary	the usage of fentanyl

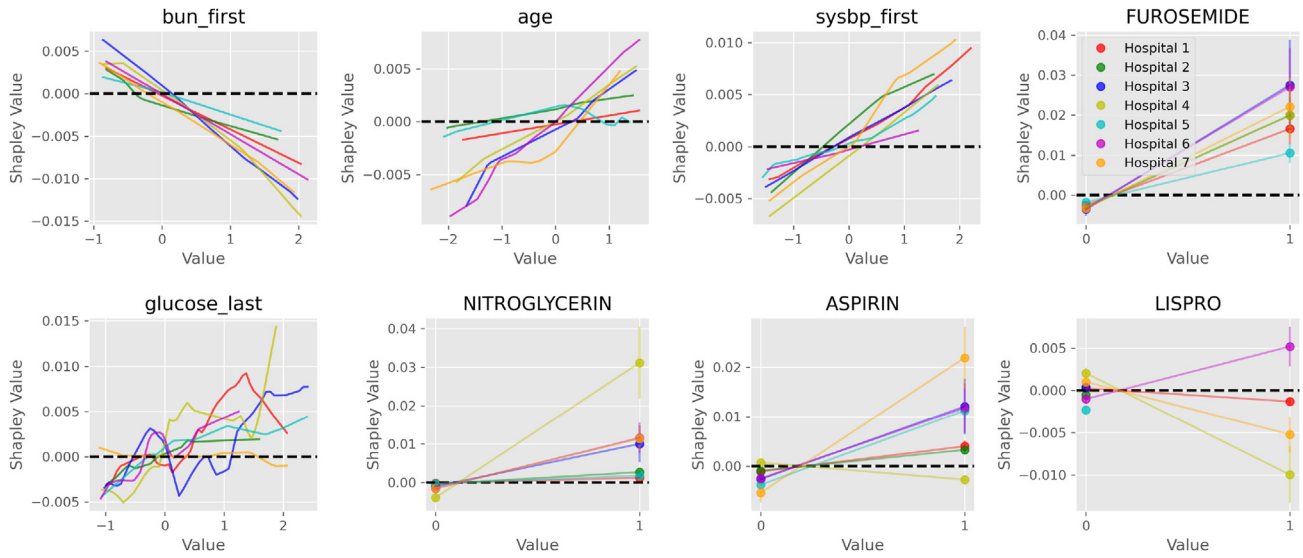

Figure 3. Shapley value plots of the top important variables for AKI prediction

The top row displays the plots of the stable features, while the down row corresponds to the plots of the specific features. Within each subfigure, the horizontal axis represents the feature value (0 or 1 for binary features, *Z* score normalized value for continuous features) and the vertical axis illustrates the Shapley values.

separation, **S** is passed into a shared network $\mathcal{G}$ while **D** is passed into a domain-specific network $\mathcal{L}^m$ where m denotes the index of the site. The outputs from $\mathcal{G}$ and $\mathcal{L}^m$ are then combined to get the final prediction $\hat{Y}$. $\mathcal{G}$ and $\{\mathcal{L}^m\}$ are implemented as multilayer perceptron (MLP). **C** does not contribute to the prediction, as it is not predictive of the outcome.

Training process

The primary objective function for training our model is the prediction loss, which is formulated as follows:

$$L_{pred} = \sum_{m=1}^{M} \frac{1}{N_m} \sum_{n=1}^{N_m} \mathrm{BCE}(y_n^m, \hat{y}_n^m),$$

where BCE is the binary cross-entropy loss and $\hat{y}_n^m$ is calculated as follows:

$$\hat{y}_n^m = \sigma(\mathcal{G}(\mathbf{s}_n^m) + \mathcal{H}^m(\mathbf{d}_n^m)),$$

where σ is the sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$.

Beside L_{pred}, we also design other objective functions to ensure the properties of **S**, **D**, and **C**: (1) Since the features in **C** are not predictive to the outcome, **S** and **D** should be the minimal feature set to build the optimal model for predicting Y. And incorporating any features in **C** will not improve the prediction performance. (2) Regarding the stable features **S**, their relationships with Y should be the same across different sites: $P^1(Y|\mathbf{S}) = \ldots = P^M(Y|\mathbf{S}) = P(Y|\mathbf{S})$ where $P^m(Y|\mathbf{S})$ is the conditional distribution of Y given **S** in the m-th site

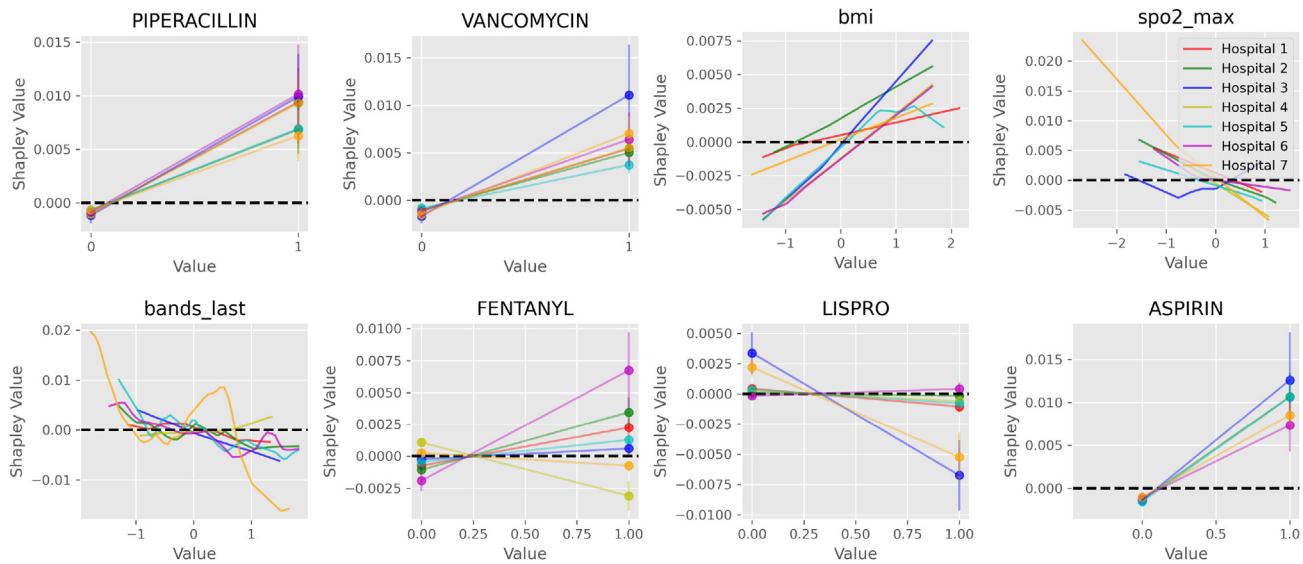

Figure 4. Shapley value plots of the top important variables for sepsis prediction

The top row displays the plots of the stable features, while the down row corresponds to the plots of the specific features. Within each subfigure, the horizontal axis represents the feature value (0 or 1 for binary features, *Z* score normalized value for continuous features) and the vertical axis illustrates the Shapley values.

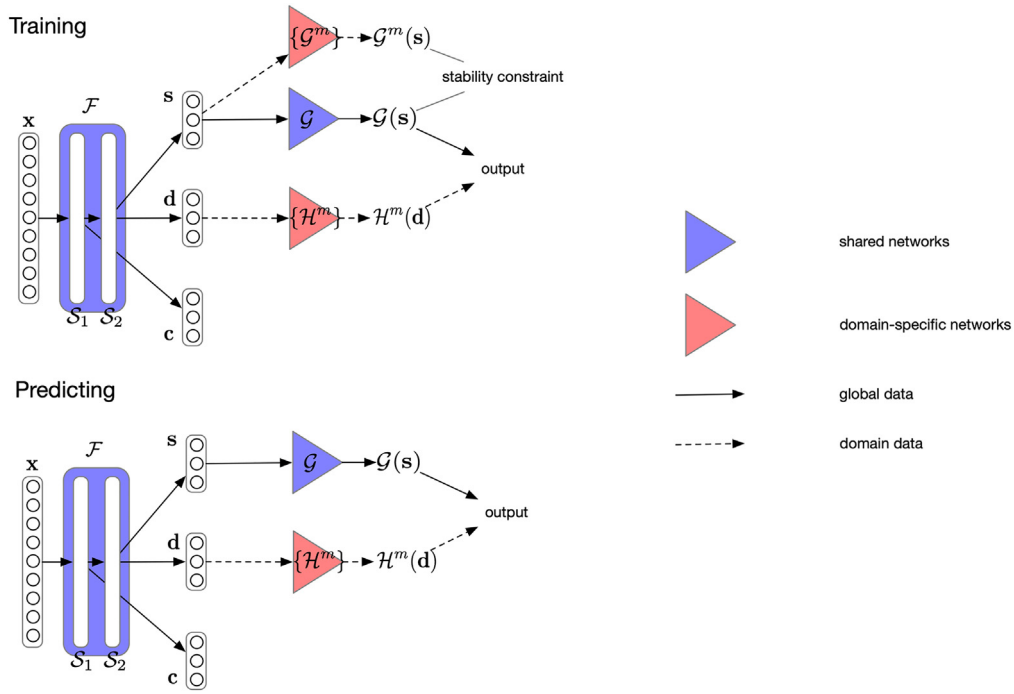

Figure 5. The structure of our proposed model

The blue components represent shared elements across different domains (sites), while the red components are domain specific. Solid arrows indicate the flow of data from all domains passing through the shared components, while dotted arrows signify data flow from each specific domain into the corresponding specific components. $\mathcal{F}$ is for separating the input features into stable, domain-specific and conditional-irrelevant parts. $\mathcal{G}$ and $\{\mathcal{H}^m\}$ are the shared and specific prediction networks, respectively, taking stable and domain-specific features as inputs. $\{\mathcal{G}^m\}$ represent the auxiliary prediction networks that cooperate with $\mathcal{G}$ to identify the stable features.

and $P(Y|\mathbf{S})$ is the conditional distribution over all sites. By considering these two properties with, we can first identify **C** by $\mathcal{S}_1$ and not include them in the prediction networks. Then we can identify **S** by $\mathcal{S}_2$, the rest are domain-specific features **D.**

For (1), we add a regularization loss to minimize the proportion of the features in **S** and **D:**

$$L_{reg} = \frac{|\mathbf{S}|+|\mathbf{D}|}{K},$$

where K is the number of the features in **X** and $|\mathbf{S}| + |\mathbf{D}|$ can be computed with the parameters of $\mathcal{S}_1$.

To ensure **S** to meet the property in (2), we need to estimate $P(Y|\mathbf{S})$ and $\{P^m(Y|\mathbf{S})\}$. we directly use $\mathcal{G}$ to estimate $P(Y|\mathbf{S})$. And we build a set of auxiliary predictors $\{\mathcal{G}^m\}$ to estimate $\{P^m(Y|\mathbf{S})\}$ where $\{\mathcal{G}^m\}$ are also implemented as MLP. With $\mathcal{G}$ and $\{\mathcal{G}^m\}$, we utilize the following loss functions with respect to the stable features:

$$L_{disc} = \sum_{m=1}^{M} \frac{1}{N_m} \sum_{n=1}^{N_m} \mathrm{DIST}\left(\sigma\left(\mathcal{G}\left(\mathbf{s}_n^m\right)\right), \sigma\left(\mathcal{G}^m\left(\mathbf{s}_n^m\right)\right)\right),$$

where DIST is the distance function and:

$$L_{grad} = \sum_{m=1}^{M} \left\| \nabla_{\theta_{\mathcal{G}}} \left(\frac{1}{N_m} \sum_{n=1}^{N_m} \mathrm{BCE}\left(y_n^m, \sigma\left(\mathcal{G}\left(s_n^m\right)\right)\right) \right) \right\|_2,$$

where $\nabla_{\theta_{\mathcal{G}}}$ is the gradient of the parameters in $\mathcal{G}$. Then the stable loss L_{stable} is defined as the summation of L_{disc} and L_{grad}.

$\mathcal{F}$, $\mathcal{G}$, and $\{\mathcal{H}^m\}$ are jointly trained to minimize the final objective function:

$$L = L_{pred} + \lambda_1 L_{stable} + \lambda_2 L_{reg},$$

where λ_1 and λ_2 are hyper-parameters to be tuned on the validation set.

The implementation details can be found in Supplementary Appendixes B and C. Supplementary Appendix B introduces the details of the optimization, while Supplementary Appendix C further introduces how to implement the learning algorithm under federated setting to avoid directly sharing the data.

Hyperparameter optimization

There are four hyperparameters in configuration: learning rate, weight decay, λ_1, and λ_2 as the weights of stable and regularization losses. A grid search was used to perform hyperparameter optimization. The hyperparameters selected are the same for both tasks as follows: learning rate = 0.01, weight decay = 0.0001, λ_1 = 0.2, and λ_2 = 0.1.

Evaluation and metrics

For each task, we established training, validation, and testing subsets through a stratified random split of 70:15:15. We use the AUROC for assessing the performance of each method across all tasks. Beside the AUROC on each site, we also reported the micro/macro AUROC over all the sites. These metrics have been used to evaluate the prediction performance across different ICU units.[41] The micro AUROC is the AUROC calculated after pooling all predictions across different sites together. And the macro AUROC is the average of the AUROC across the sites. For individual AUROC(s) calculated at hospitals and macro/micro AUROC, we conducted 10 runs and reported the average results.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2023.100898.

ACKNOWLEDGMENTS

R.S. would like to acknowledge the support from Tri-Institutional Training Program in Computational Biology and Medicine (CBM) funded by the NIH grant 1T32GM083937. Z.X., W.P., and F.W. would like to acknowledge the support from NSF 1750326, NSF 2212175, NIH R01AG076234, NIH RF1AG072449, NIH R01MH124740, NIH R01AG080624, NIH RF1AG084178, R01AG080991, Google Faculty Research Award, and Amazon Machine Learning Research Award.

AUTHOR CONTRIBUTIONS

Conceptualization, W.P. and F.W.; methodology, W.P. and F.W.; investigation, W.P., Z.X., and F.W.; writing – original draft, W.P., Z.X., and F.W.; writing – review and editing, W.P., Z.X., R.S., and F.W.; supervision, F.W.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 29, 2023
Revised: September 6, 2023
Accepted: November 21, 2023
Published: December 26, 2023

REFERENCES

1. Shickel, B., Tighe, P.J., Bihorac, A., and Rashidi, P. (2018). Deep EHR: a survey of recent advances in deep learning techniques for electronic healthIcord (EHR) analysis. IEEE J. Biomed. Health Inform. *22*, 1589–1604.
2. Yang, J., Soltan, A.A.S., and Clifton, D.A. (2022). Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. NPJ Digit. Med. *5*, 69.
3. Singh, H., Mhasawade, V., and Chunara, R. (2022). Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. PLOS Digit. Health *1*, e0000023.
4. Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R.R., and Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. *10*, 12598.
5. Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Trans. Intell. Syst. Technol. *10*, 1–19.
6. Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., Liu, A., Costa, A.B., Wood, B.J., Tsai, C.S., et al. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. Nat. Med. *27*, 1735–1743.
7. Vaid, A., Jaladanki, S.K., Xu, J., Teng, S., Kumar, A., Lee, S., Somani, S., Paranjpe, I., De Freitas, J.K., Wanyan, T., et al. (2021). Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: Machine learning approach. JMIR Med. Inform. *9*, e24207.
8. Song, X., Yu, A.S.L., Kellum, J.A., Waitman, L.R., Matheny, M.E., Simpson, S.Q., Hu, Y., and Liu, M. (2020). Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nat. Commun. *11*, 5668.
9. Rajendran, S., Xu, Z., Pan, W., Ghosh, A., and Wang, F. (2023). Data heterogeneity in federated learning with Electronic Health Records: Case studies of risk prediction for acute kidney injury and sepsis diseases in critical care. PLOS Digit. Health *2*, e0000117.
10. Xing, H., Xiao, Z., Qu, R., Zhu, Z., and Zhao, B. (2022). An efficient federated distillation learning system for Multitask Time Series classification. IEEE Trans. Instrum. Meas. *71*, 1–12.
11. Crowson, M.G., Moukheiber, D., Arévalo, A.R., Lam, B.D., Mantena, S., Rana, A., Goss, D., Bates, D.W., and Celi, L.A. (2022). A systematic review of federated learning applications for biomedical data. PLOS Digit. Health *1*, e0000033.
12. Wang, H., Zhang, X., Xia, Y., and Wu, X. (2023). An intelligent blockchain-based access control framework with federated learning for genome-wide association studies. Comput. Stand. Interfac. *84*, 103694.
13. Wu, X., Zhang, Y., Shi, M., Li, P., Li, R., and Xiong, N.N. (2022). An adaptive federated learning scheme with differential privacy preserving. Future Generat. Comput. Syst. *127*, 362–372.
14. Pollard, T.J., Johnson, A.E.W., Raffa, J.D., Celi, L.A., Mark, R.G., and Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci. Data *5*, 180178.
15. Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.D., Coopersmith, C.M., et al. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA *315*, 801–810.
16. Khwaja, A. (2012). KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin. Pract. *120*, c179–c184.
17. Nemati, S., Holder, A., Razmi, F., Stanley, M.D., Clifford, G.D., and Buchman, T.G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit. Care Med. *46*, 547–553.
18. Xu, Z., Feng, Y., Li, Y., Srivastava, A., Adekkanattu, P., Ancker, J.S., Jiang, G., Kiefer, R.C., Lee, K., Pacheco, J.A., et al. (2019). Predictive modeling of the risk of acute kidney injury in critical care: a systematic investigation of the class imbalance problem. AMIA Jt. Summits Transl. Sci. Proc. *2019*, 809–818.
19. Duong, L., Cohn, T., Bird, S., and Cook, P. (2015). Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing, pp. 845–850.
20. Evgeniou, T., and Pontil, M. (2004). Regularized multi–task learning. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 109–117.
21. Liu, P., Qiu, X., and Huang, X. (2017). Adversarial multi-task learning for text classification. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1–10.
22. Luo, J., and Wu, S. (2022). Adapt to adaptation: Learning personalization for cross-silo federated learning. In IJCAI: proceedings of the conference, p. 2166.
23. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: uniform manifold approximation and projection. J. Open Source Softw. *3*, 861.
24. Lundberg, S.M., and Lee, S.I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777.
25. Qian, Q., Wu, J., Wang, J., Sun, H., and Yang, L. (2021). Prediction models for AKI in ICU: a comparative study. Int. J. Gen. Med. *14*, 623–632.
26. Vincent, J.L., Sakr, Y., Singer, M., Martin-Loeches, I., Machado, F.R., Marshall, J.C., Finfer, S., Pelosi, P., Brazzi, L., Aditianingsih, D., et al. (2020). Prevalence and outcomes of infection among patients in intensive care units in 2017. JAMA *323*, 1478–1487.
27. He, Z., Du, L., Zhang, P., Zhao, R., Chen, X., and Fang, Z. (2020). Early sepsis prediction using ensemble learning with deep features and artificial features extracted from clinical electronic health records. Crit. Care Med. *48*, e1337–e1342.
28. Liu, Y.L., Prowle, J., Licari, E., Uchino, S., and Bellomo, R. (2009). Changes in blood pressure before the development of nosocomial acute kidney injury. Nephrol. Dial. Transplant. *24*, 504–511.
29. Uchino, S., Bellomo, R., and Goldsmith, D. (2012). The meaning of the blood urea nitrogen/creatinine ratio in acute kidney injury. Clin. Kidney J. *5*, 187–191.
30. Okubo, K., Kurosawa, M., Kamiya, M., Urano, Y., Suzuki, A., Yamamoto, K., Hase, K., Homma, K., Sasaki, J., Miyauchi, H., et al. (2018). Macrophage extracellular trap formation promoted by platelet activation is a key mediator of rhabdomyolysis-induced acute kidney injury. Nat. Med. *24*, 232–238.

31. Koyner, J.L., Davison, D.L., Brasha-Mitchell, E., Chalikonda, D.M., Arthur, J.M., Shaw, A.D., Tumlin, J.A., Trevino, S.A., Bennett, M.R., Kimmel, P.L., et al. (2015). Furosemide stress test and biomarkers for the prediction of AKI severity. J. Am. Soc. Nephrol. *26*, 2023–2031.

32. Wang, H.E., Griffin, R., Judd, S., Shapiro, N.I., and Safford, M.M. (2013). Obesity and risk of sepsis: A population-based cohort study. Obesity *21*, E762–E769.

33. Chaturvedi, R., Burton, B.N., Trivedi, S., Schmidt, U.H., and Gabriel, R.A. (2022). The Association of preoperative hematocrit with adverse events following exploratory laparotomy in septic patients: a retrospective analysis. J. Intensive Care Med. *37*, 46–51.

34. De Freitas, G., Gudur, A., Vela-Ortiz, M., Jodelka, J., Livert, D., and Krishnamurthy, M. (2019). Where there is sodium there may be sepsis. J. Community Hosp. Intern. Med. Perspect. *9*, 296–299.

35. Shashikumar, S.P., Stanley, M.D., Sadiq, I., Li, Q., Holder, A., Clifford, G.D., and Nemati, S. (2017). Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. J. Electrocardiol. *50*, 739–743.

36. Pollard, T., Johnson, A., Raffa, J., Celi, L.A., Badawi, O., and Mark, R. (2019). eICU Collaborative Research Database (PhysioNet) version 2.0. (PhysioNet). https://doi.org/10.13026/C2WM1R.

37. Pan, W., Xu, Z.C., Rajendran, S., and Wang, F. (2023). Code, Datasets, and Results for the Paper "An Adaptive Federated Learning Framework for Clinical Risk Prediction with Electronic Health Records from Multiple Hospitals." (Zenodo). https://doi.org/10.5281/zenodo.10080736.

38. Reyna, M.A., Josef, C.S., Jeter, R., Shashikumar, S.P., Westover, M.B., Nemati, S., Clifford, G.D., and Sharma, A. (2020). Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Crit. Care Med. *48*, 210–217.

39. Wang, R.Z., Sun, C.H., Schroeder, P.H., Ameko, M.K., Moore, C.C., and Barnes, L.E. (2018). Predictive models of sepsis in adult ICU patients. In IEEE International Conference on Healthcare Informatics (ICHI), pp. 390–391.

40. Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. (2020). Feature selection using stochastic gates. In International Conference on Machine Learning, pp. 10648–10659.

41. Suresh, H., Gong, J.J., and Guttag, J.V. (2018). Learning tasks for multi-task learning: Heterogenous patient populations in the icu. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 802–810.