PLOS DIGITAL HEALTH

RESEARCH ARTICLE

Data heterogeneity in federated learning with Electronic Health Records: Case studies of risk prediction for acute kidney injury and sepsis diseases in critical care

Suraj Rajendran[1], **Zhenxing Xu**[2], **Weishen Pan**[2], **Arnab Ghosh**[3], **Fei Wang**[2]*

1 Tri-Institutional Computational Biology & Medicine Program, Cornell University, New York, New York, United States of America, **2** Division of Health Informatics, Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, United States of America, **3** Departments of Medicine, Weill Cornell Medical College, Cornell University, New York, New York, United States of America

* few2001@med.cornell.edu

Citation: Rajendran S, Xu Z, Pan W, Ghosh A, 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(3): e0000117. https://doi.org/10.1371/journal.pdig.0000117

Editor: Martin G. Frasch, University of Washington, UNITED STATES

Received: September 1, 2022

Accepted: February 10, 2023

Published: March 15, 2023

Data Availability Statement: Patient data was extracted from the eICU Collaborative Research Database, a multi-center critical care database made publicly available through Philips Healthcare and the MIT Laboratory for Computational Physiology (https://eicu-crd.mit.edu/). Processed data and scripts used for analyses are also available at https://github.com/surajraj99/Data-Heterogeneity-in-Federated-Learning.

Abstract

With the wider availability of healthcare data such as Electronic Health Records (EHR), more and more data-driven based approaches have been proposed to improve the quality-of-care delivery. Predictive modeling, which aims at building computational models for predicting clinical risk, is a popular research topic in healthcare analytics. However, concerns about privacy of healthcare data may hinder the development of effective predictive models that are generalizable because this often requires rich diverse data from multiple clinical institutions. Recently, federated learning (FL) has demonstrated promise in addressing this concern. However, data heterogeneity from different local participating sites may affect prediction performance of federated models. Due to acute kidney injury (AKI) and sepsis' high prevalence among patients admitted to intensive care units (ICU), the early prediction of these conditions based on AI is an important topic in critical care medicine. In this study, we take AKI and sepsis onset risk prediction in ICU as two examples to explore the impact of data heterogeneity in the FL framework as well as compare performances across frameworks. We built predictive models based on local, pooled, and FL frameworks using EHR data across multiple hospitals. The local framework only used data from each site itself. The pooled framework combined data from all sites. In the FL framework, each local site did not have access to other sites' data. A model was updated locally, and its parameters were shared to a central aggregator, which was used to update the federated model's parameters and then subsequently, shared with each site. We found models built within a FL framework outperformed local counterparts. Then, we analyzed variable importance discrepancies across sites and frameworks. Finally, we explored potential sources of the heterogeneity within the EHR data. The different distributions of demographic profiles, medication use, and site information contributed to data heterogeneity.

Funding: SR would like to acknowledge the support from Tri-Institutional Training Program in Computational Biology and Medicine (CBM) funded by the NIH grant 1T32GM083937. ZX, WP, FW would like to acknowledge the support from NSF 1750326, NSF 2212175, NIH R01AG076234, NIH RF1AG072449, Google Faculty Research Award and Amazon Machine Learning Research Award. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

The availability of a large amount of healthcare data such as Electronic Health Records (EHR) and advances of artificial intelligence (AI) techniques provides opportunities to build predictive models for disease risk prediction. Due to the sensitive nature of healthcare data, it is challenging to collect the data together from different hospitals and train a unified model on the combined data. Recent federated learning (FL) demonstrates promise in addressing the fragmented healthcare data sources with privacy-preservation. However, data heterogeneity in the FL framework may influence prediction performance. Exploring the heterogeneity of data sources would contribute to building accurate disease risk prediction models in FL. In this study, we take acute kidney injury (AKI) and sepsis prediction in intensive care units (ICU) as two examples to explore the effects of data heterogeneity in the FL framework for disease risk prediction using EHR data across multiple hospital sites. In particular, multiple predictive models were built based on local, pooled, and FL frameworks. The local framework only used data from each site itself. The pooled framework combined data from all sites. In the FL framework, each local site did not have access to other sites' data. We found models built within a FL framework outperformed local counterparts. Then, we analyzed variable importance discrepancies across sites and frameworks. Finally, we explored potential sources of the heterogeneity within EHR data. The different distributions of demographic profiles, medication use, site information such as the type of ICU at admission contributed to data heterogeneity.

Introduction

Acute kidney injury (AKI) and sepsis are two types of potentially life-threatening clinical conditions that complicate treatment, clinical trajectories, and potentially worsen outcomes of a significant number of hospitalized or intensive care unit (ICU)-patients [1–2]. For patients with AKI or sepsis, morbidity and mortality are usually higher than patients without AKI or sepsis, with as much as a sevenfold increased mortality risk, regardless of type of ICU (for example, medical, surgical, or cardiac) [3–4]. Moreover, healthcare utilization within the ICU is often higher for patients with these conditions. For example, patients with AKI and sepsis often require hemodialysis, inotropic support, or mechanical ventilation [5]. Therefore, early prediction of AKI or sepsis risk in critical care settings can facilitate early interventions that are likely to provide benefit, including aggressive treatment with fluid resuscitation and antimicrobials that may improve patient outcomes [6].

Recently, due to wider availability of electronic health record (EHR) data and advances in artificial intelligence (AI), machine learning (ML) based disease risk prediction has attracted more attention in the ICU setting [7]. Previous studies on AKI and sepsis onset risk prediction mainly focused on building a predictive model on medical data from single hospitals [8–13]. However, building an accurate and generalizable disease risk prediction model requires a large amount of data from a diverse patient population [8]. Collecting the data together from different hospitals and constructing a unified risk prediction model on the combined data can lead to better prediction performance. Moreover, using multiple hospitals or sites data over single institution data can add to the generalizability of ML models [14]. A recent study has shown that creating more generalizable models can increase algorithmic fairness, yet many published models lack this generalizability across geographic locations and demographics [15]. However, due to the highly sensitive nature of EHR in terms of protected health information (PHI) of patients, aggregating multiple institutions' data all together is challenging [16].

More recently, federated learning (FL) has emerged as a promising strategy on building ML models with fragmented sensitive data [17]. FL is one mechanism of training ML models across multiple decentralized sites holding local data samples without exchanging them [18]. It builds a central aggregator to obtain global ML model's parameters by iteratively exchanging model parameters from local ML models. However, data heterogeneity in the FL framework may affect prediction performance [19]. For example, different hospitals have different populations, which may have a high degree of variability in the patient treatment, such as different medications they administer and different procedures they conduct. This heterogeneity especially affects the performance of sepsis and AKI prediction models which rely on patient demographics, disease history, and medications [20]. Both AKI and sepsis are also highly heterogeneous [21]. This makes models built with conventional FL strategies such as federated averaging challenging to generalize across clinics, limiting their use [7,22,23]. Several federated architectures have been proposed to mitigate effects of data heterogeneity in other domains and built personalized, but globally correlated, models to mitigate drift across sites [23], such as model-agnostic meta-learning (MAML), federated multitask learning, and knowledge distillation [24–28]. However, it is not clear how such data heterogeneity problem will impact building risk prediction models in clinical medicine.

To fill this research gap, we comprehensively investigate the effects of data heterogeneity in the FL framework for predicting the onset risk of AKI and sepsis in ICU setting using EHR data from multiple hospital sites. We built multiple predictive models in local, pooled, and FL settings. The local setting built an individual model for each site from its own data. The pooled setting built a global model shared across all sites with their combined data. The FL setting also built a global model, where each local site did not share data with others, but updated model parameters locally and shared the updated model parameters to a central aggregator, which was used to update the global model parameters and shared back with each site. By comparing the performance of models trained from different settings with each other, we investigated how data heterogeneity would impact the federated risk prediction models. We also explored the potential sources of the heterogeneity within EHR data by analyzing predictor importance across settings and sites. The differences were contrasted according to patient and hospital information to elucidate sources of heterogeneity and how they would potentially impact the different predictive modeling settings. The overall workflow of our study is shown in Fig 1.

The notable contributions of this work to the literature are as follows:

- With the context of AKI and sepsis onset risk prediction in ICU setting, a comprehensive comparison in terms of prediction performance among local, pooled, and federated settings were conducted with a set of ML models.
- We have identified important predictors for AKI and sepsis risk and performed exhaustive analysis on they would impact the prediction results. These predictors can be used by medical specialists to monitor the risk of AKI and sepsis for patients in ICU, while accounting for the specifics of their own hospitals. In addition, we have delineated differences in feature importance across medical sites, outlining metrics for direct comparison of feature importance across different settings (i.e., local, pooled, and federated).
- We have performed a thorough analysis on the potential sources of heterogeneity between hospital sites according to patient demographic, medication, and lab data, as well as hospital information such as available unit types. We outline how these sources of heterogeneity could be connected to the varying predictor importance derived across sites and settings.

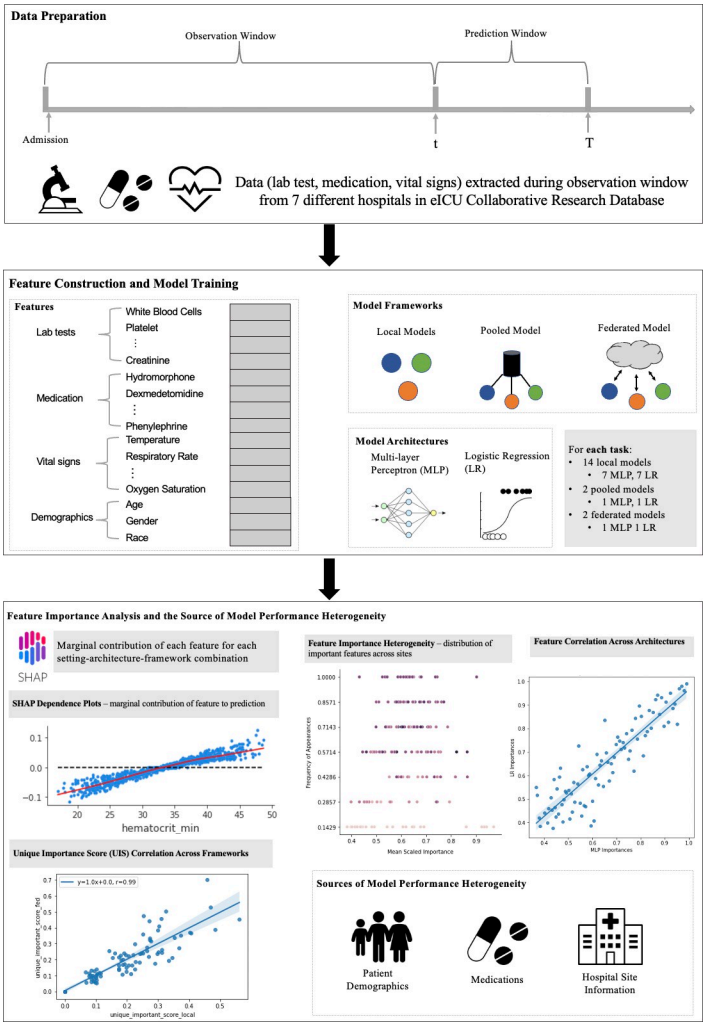

Fig 1. The framework of the study. In Data Preparation, different types of data including lab test, medication, vital signs, and demographic are extracted during the observation window, which are used to build patients' profiles to predict whether they would suffer from acute kidney injury or sepsis in the prediction window. In Feature Construction and Model Training, individual features from lab test, medication, vital signs, and demographic were obtained to build a predictive model based on three frameworks including local, pooled, and federated frameworks. In each framework, two common model architectures including logistic regression and multi-layer perceptron were used. In Feature Importance Analysis and Source of Model Performance Heterogeneity, feature importance heterogeneity, feature correlation across mode architectures and frameworks, and sources of model performance heterogeneity were explored.

https://doi.org/10.1371/journal.pdig.0000117.g001

Results

Development of AKI and sepsis prediction models in local sites

Data for 21,796 AKI patients and 22,0082 sepsis patients at 7 hospitals were extracted from the eICU collaborative research database, following inclusion and exclusion criteria denoted in the "Methods" section. All patients shared 354 unique variables which included lab tests, vital signs, demographics, and medications. AKI patients were labeled both within a 24h and 48h observation window, leading to two settings for AKI prediction. For sepsis, we labeled patient data in accordance with Sepsis-3 clinical criteria. We predicted whether patients would suffer from sepsis 6 hours prior to onset, onset point included. Within the observation window, lab

tests and vital sign information were aggregated through several statistics (minimum, maximum, first, and last values) into several new features. Three model frameworks were designed including local, pooled, and federated model architectures. The details of each model architecture were described in the "Methods" section. For each framework, two model architectures including the multilayer perceptron (MLP) and logistic regression (LR) were explored. Processed data and scripts used for analyses are also available at https://github.com/surajraj99/Data-Heterogeneity-in-Federated-Learning. On the Apple M1 Max with 10-core CPU, local LR models were trained within an average of 2.54 ± 0.84 seconds and local MLP models were trained in 2.36 ± 0.79 seconds. The large standard deviation of local performances is due to varying dataset sizes across different sites. Pooled LR and MLP models were trained within 18.94 ± 0.05 and 19.33 ± 0.06 seconds, respectively. Federated LR and MLP models were trained within 51.63 ± 24.26 and 58.51 ± 27.45 seconds, respectively. Communication costs between sites and the central sever within the federated framework take a significant amount of time, as evidenced by the time differences between pooled and federated framework. These communication costs can be reduced by increasing the number of epochs that local sites train models at each federated framework.

Fig 2 illustrates the LR and MLP performance, measured by area-under-receiver-operator-curve (AUC), on both AKI 24h and 48h settings. Sepsis prediction setting results can be found in S2 Text and S6 Fig. We observed:

- When using local model framework: AKI 24h LR models performed within the range of 0.680–0.809, whereas MLP models performed within a range of 0.677–0.821. Similarly, AKI 48h LR models performed within the range of 0.680–0.809, whereas MLP models performed within a range of 0.673–0.800. Sepsis LR models' performances ranged between 0.771–0.834 across sites, whereas MLP models' performances ranged between 0.772–0.829. The LR and MLP models performed similarly across all prediction tasks.
- When using pooled model framework: AKI 24h LR models performed within the range of 0.672–0.742, whereas MLP models performed within a range of 0.78–0.827. Similarly, pooled AKI 48h LR models performed within the range of 0.683–0.744, whereas MLP models performed within a range of 0.686–0.755. Pooled sepsis LR models' performances ranged between 0.731–0.800 across sites, whereas MLP models' performances ranged between 0.732–0.793. LR pooled models showed more consistent performances to local model counterparts in comparison to MLP pooled models.
- When using federated model framework: AKI 24h LR models performed within the range of 0.742–0.834, whereas MLP models performed within a range of 0.732–0.839. Similarly, AKI 48h LR models performed within the range of 0.722–0.835, whereas MLP models performed within a range of 0.72–0.833. Federated sepsis LR models' performances ranged between 0.833–0.862 across sites, whereas MLP models' performances ranged between 0.823–0.861.
- Generally, the federated model outperformed the local model and pooled model. The pooled models underperformed the local model.

Clinical interpretation of sepsis and AKI prediction models

Using Shapley Additive exPlanations (SHAP) values, we investigated the marginal effects of the features identified as predictive by each model. Fig 3 illustrates the marginal plots (SHAP dependence plots) for the top 10 most important features for each pooled model on the AKI prediction settings. Fig 4 shows the SHAP dependence plots for AKI setting federated models. Dependence plots for all local models are available in S4 Fig. Sepsis prediction results are

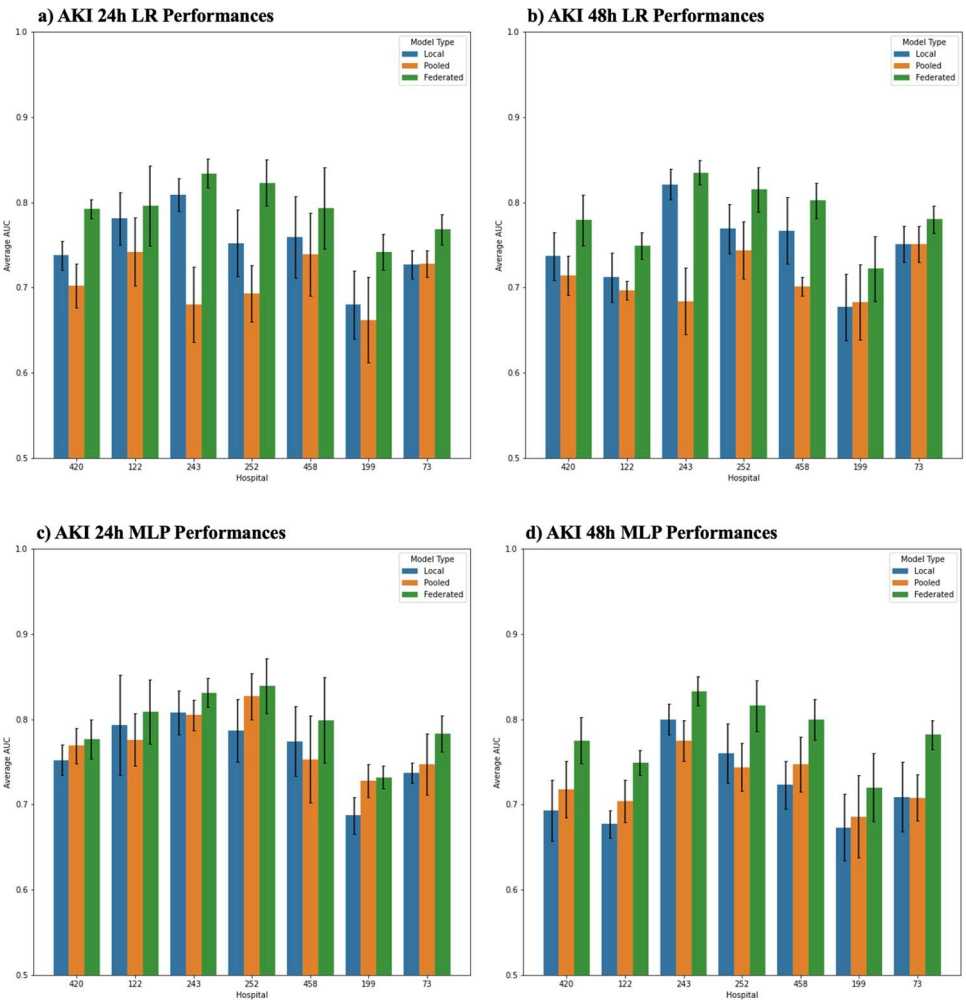

Fig 2. Area Under Receiver Operating Curve (AUROC) for AKI Setting. Each plot shows performances for AKI 24h and 48h prediction settings. Blue bars depict each local site's model performance on their respective site test data. Orange bars depict pooled model performance on each local site's test data. Green bars depict federated model performance on each local site's test data.

https://doi.org/10.1371/journal.pdig.0000117.g002

available in the Supplemental Information and S7 Fig. To calculate the SHAP values for 1000 samples on the Apple M1 Max with 10-core CPU took on average 1076 ± 52 seconds.

In the AKI 24h setting, the pooled MLP model identified last measured level of creatinine *(creatinine_last)*, last measured hematocrit level *(hematocrit_last)*, Furosemide, *bg_paco2_min*, maximum potassium level *(potassium_max)*, minimum creatinine levels *(creatinine_min)*, last measured systolic blood pressure *(sysbp_last)*, *hemoglobin_first*, minimum bicarbonate level *(bicarbonate_min)*, and last measured calcium level *(calcium_last)* as the top 10 most important variables. All factors except Furosemide were lab tests and vital signs. The pooled LR model shared several important factors with the pooled MLP model, with the addition of age, first measured calcium level *(calcium_first)*, and last measured blood urea nitrogen level *(bun_last)*. Of note, in the pooled MLP model, *creatinine_last* of ~4 mg/dL is associated with an exp(0.4) = 1.5-fold increase in risk of AKI 24h. In the pooled LR model, *creatinine_last* shows a similarly strong relationship as the pooled MLP to AKI 24h risk. A *bun_last* measurement of ~60 mg/dL is associated with a exp(0.2) = 1.2-fold increase in risk of AKI 24h. In the

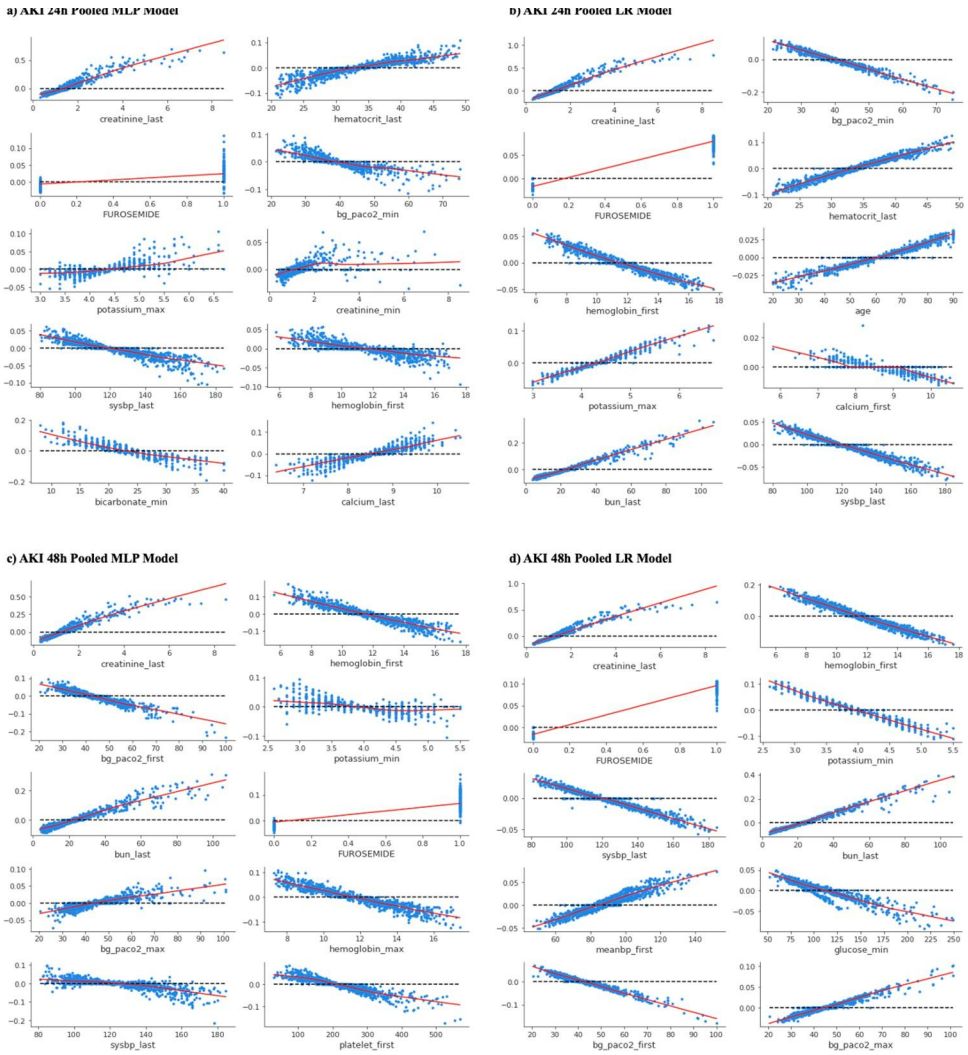

Fig 3. Shapley dependence plots for top 10 features for pooled models. Each panel shows the marginal effects of each of the most impactful features ranked among the top 10 for predicting AKI 24h or 48h using pooled models. The x-axis gives the raw values of each feature, and the y-axis gives the logarithmic of estimated odds ratio (i.e., the SHAP value) for sepsis, AKI 24h or AKI 48h, when a feature takes a certain value. Each dot represents the SHAP value of a sample. The LOWESS curve, used for smoother extrapolating across all the dots, is plotted in red for all panels. (a, c) show Shapley dependence plots for pooled MLP models and (b, d) show Shapley dependence plots for pooled LR models. (a, b) show plots for AKI 24h, and (c, d) show plots for AKI 48h.

https://doi.org/10.1371/journal.pdig.0000117.g003

pooled LR model, the risk of AKI 24h given administration of furosemide, is greater than AKI risk in the MLP model, with an odds ratio of exp(0.1) = 1.1.

In the AKI 48h setting, the pooled MLP model identified *creatinine_last*, *hemoglobin_first*, *bg_paco2_first*, *potassium_min*, *bun_last*, Furosemide, maximum partial pressure of carbon dioxide (*bg_paco2_max*), *hemoglobin_max*, *sysbp_last*, and first measured platelet count *(platlet_first)* as the top 10 most important variables. All factors except Furosemide were lab tests and vital signs. The pooled LR model shared several important factors with the pooled model, with the addition of the mean systolic and diastolic blood pressure *(meanbp_first)* and minimum glucose level *(glucose_min)*. Like the 24h setting, in the 48h pooled MLP model, *creatinine_last* of ~4 mg/dL is associated with an exp(0.4) = 1.5-fold increase in risk of AKI. A *bun_last* measurement of greater than ~25 mg/dL is associated with an increased risk of AKI

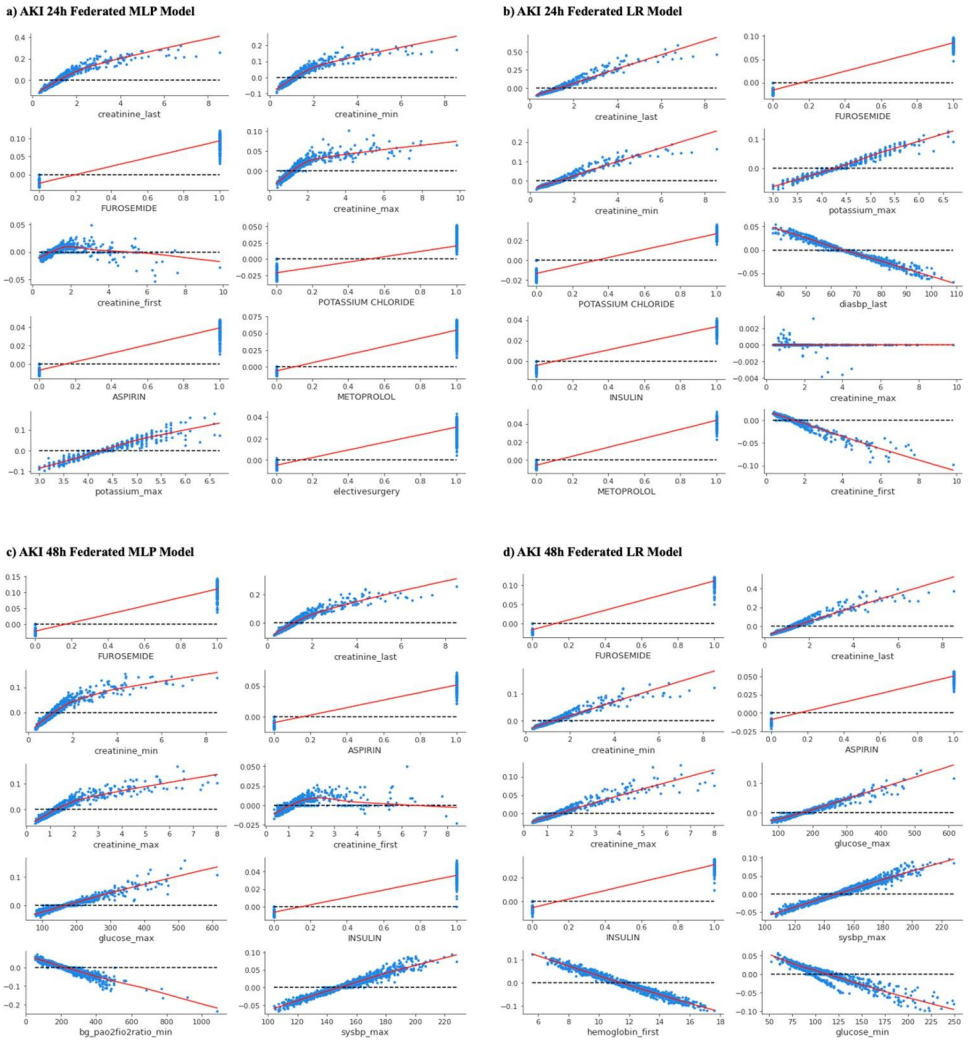

Fig 4. Shapley dependence plots for top 10 features for federated models. Each panel shows the marginal effects of each of the most impactful features ranked among the top 10 for predicting AKI 24h or 48h using federated models. The x-axis gives the raw values of each feature, and the y-axis gives the logarithmic of estimated odds ratio (i.e., the SHAP value) for sepsis, AKI 24h or AKI 48h, when a feature takes a certain value. Each dot represents the SHAP value of a sample. The LOWESS curve, used for smoother extrapolating across all the dots, is plotted in red for all panels. (a, c) show Shapley dependence plots for federated MLP models and (b, d) show Shapley dependence plots for federated LR models. (a, b) show plots for AKI 24h, and (c, d) show plots for AKI 48h.

https://doi.org/10.1371/journal.pdig.0000117.g004

48h. In the pooled LR model, *creatinine_last* and *bun_last* show similarly strong relationships as the pooled MLP model. Furosemide is considered an important medication across all AKI settings and model architectures.

For the AKI 24h setting, the federated MLP and LR model consider more medications important than their respective pooled counterparts. Medications considered important by the federated MLP model include Furosemide, Potassium Chloride, Aspirin, and Metoprolol, whereas the federated LR model considered Insulin important as well. Interestingly, the federated MLP model considered the patient's choice of elective surgery *(electivesurgery)* as an important feature, albeit a relatively small increase (exp(0.02) = 1.02-fold) in risk of AKI 24h. Like the 24h setting, the federated MLP and LR models of the AKI 48h setting considered more medications important than their respective pooled counterparts. Both the MLP and LR

model consider administration of Aspirin and Insulin as important factors. The federated MLP for the 48h setting uniquely finds the minimum ratio of "partial pressure of oxygen" to "fractional inspired oxygen" *(bg_pao2fio2ratio_min)* and maximum level of glucose *(glucose_-max)* as important factors. Local models shared numerous important factors with pooled and federated models, depicting similar relationships between feature value and risk of sepsis/AKI (S4 Fig).

Source of prediction performance heterogeneity across model architectures, frameworks, and sites

To better understand differences in feature importances across hospital sites and model frameworks, we performed a qualitative analysis which looked at the most important variables selected by models and their prevalence across sites. Fig 5, 6, and 7 show features in relation to

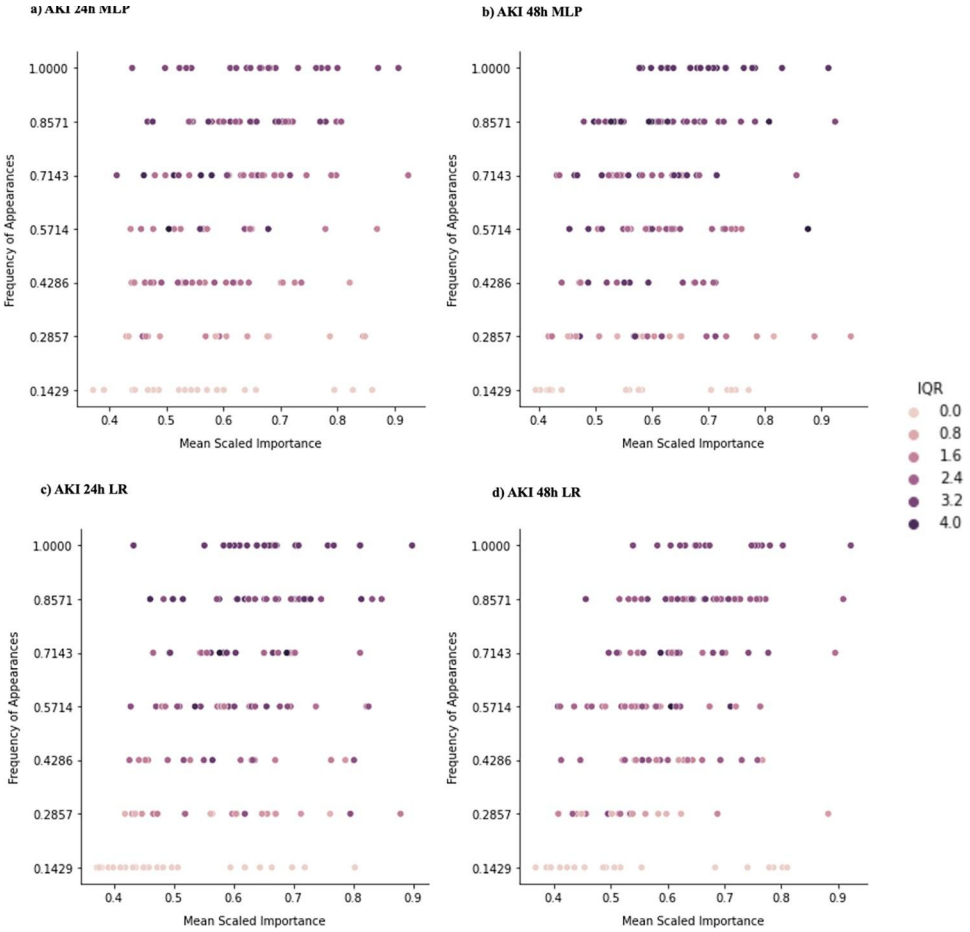

Fig 5. Distribution of important features at all local models across local sites. The figure demonstrates feature importance disparities for all AKI settings (24h, 48h) local models (MLP and LR). (a, b) show feature importance disparities for MLP models. (c, d) show feature importances for LR models. Each dot corresponds to one of the most important features ranked among the top-100 by at least one of the seven models; y-axis measures the proportions of sites that identified the feature as top-100, or "commonality across sites"; x-axis measures the mean of feature importance rankings measured as "soft ranking" (the closer it is to 1, the higher the feature ranks). Top-100 is an arbitrary cutoff we used to analyze the most important features to illustrate heterogeneity. Each feature is also color coded by the interquartile range (IQR) of the ranks across sites (the higher the IQR is, the more disagreement across sites on the importance of that feature).

https://doi.org/10.1371/journal.pdig.0000117.g005

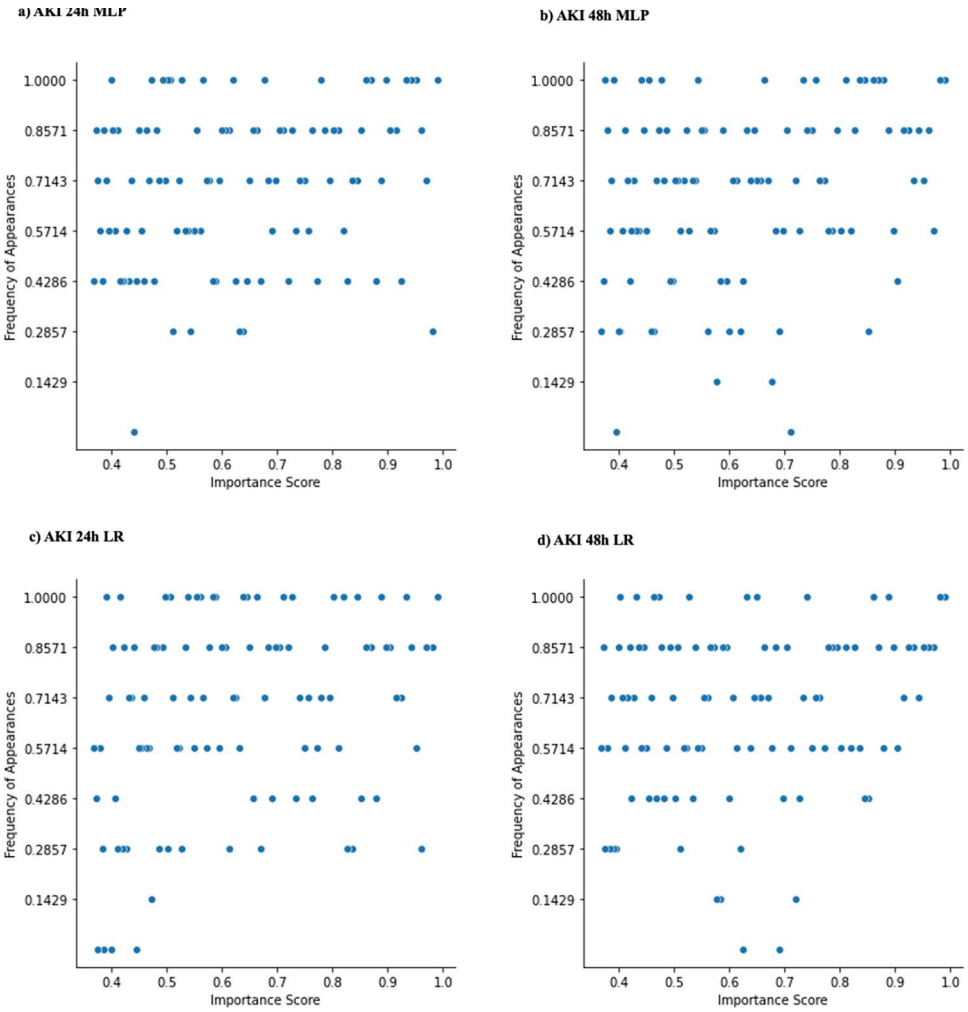

Fig 6. Distribution of important features at pooled models across local sites. The figure demonstrates feature importance disparities for all AKI settings (24h, 48h) pooled models (MLP and LR). Each dot corresponds to one of the most important features ranked among the top-100 by the pooled model; y-axis measures the commonality across sites; x-axis measures the feature importance soft rankings.

https://doi.org/10.1371/journal.pdig.0000117.g006

their importance rankings in AKI prediction models, where the y-axis is the proportion of sites that consider the feature as a top 100 feature (for the specific model architecture). For example, a feature that has a y-value of 1.0 is deemed important at all sites, whereas if a feature has a y-value of 0.1429 (1/7), it is only considered important at one site. The x-axis shows the importance ranking of the feature, averaged across the sites it is considered important (i.e., top 100) in (i.e., a feature is more important if it is closer to 1). Results for the sepsis prediction setting are available in the Supplemental Information and S8 Fig.

Fig 5 shows the distribution of important features across sites for local models. For all settings, there are features that are both 'universally important' and site-specific (i.e., important at only a subset of sites). The universally important features across most sites for AKI (both 24h and 48h) included *creatinine_last*, *creatinine_min*, *creatinine_max*, administration of Ondansteron, *glucose_max*, and *urineoutput_sum*. However, the relative importance of features at local sites were different. This disagreement was reflected in the dependence plots for the local models (S4 Fig). Universally important features like *creatinine_last*, administration of Sodium

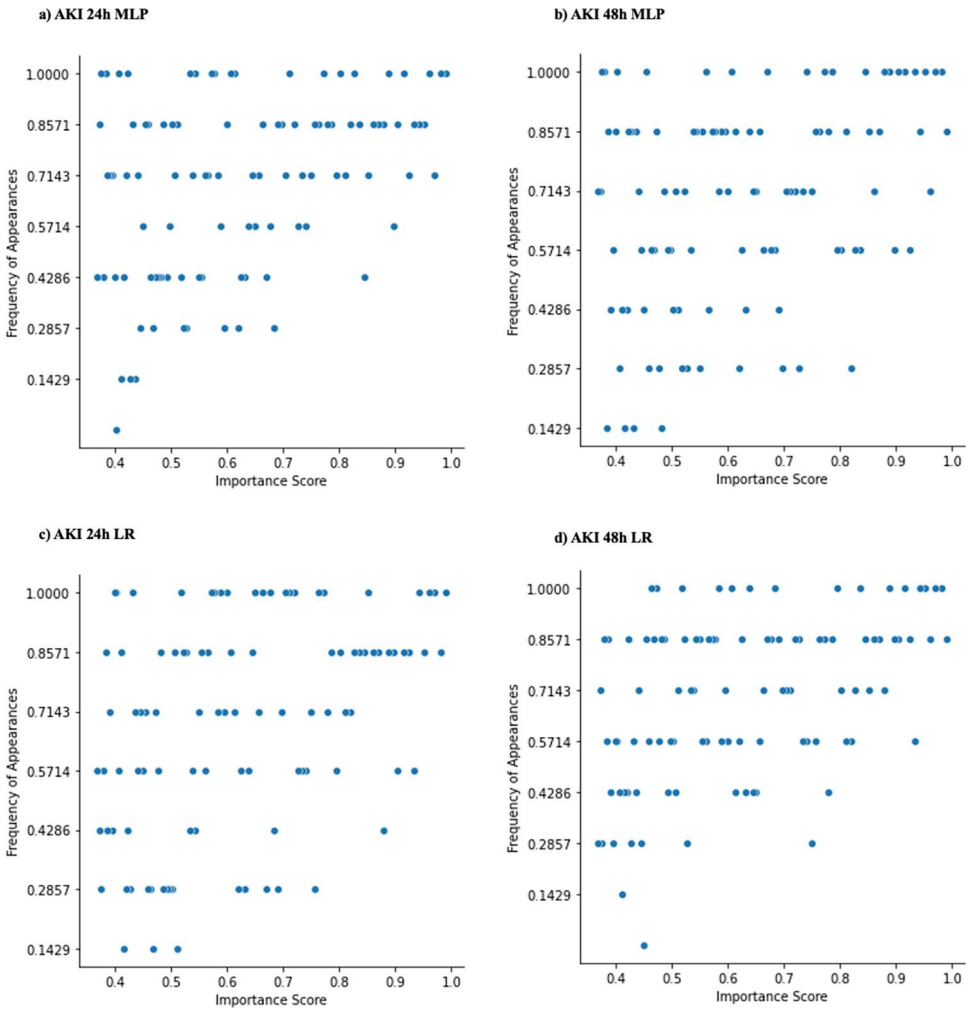

Fig 7. Distribution of important features at federated models across local sites. The figure demonstrates feature importance disparities for all AKI settings (24h, 48h) federated models (MLP and LR). Each dot corresponds to one of the most important features ranked among the top-100 by the pooled model; y-axis measures the commonality across sites; x-axis measures the feature importance soft rankings.

https://doi.org/10.1371/journal.pdig.0000117.g007

Chloride, among others, have different trends with the prediction diagnosis depending on the site.

Fig 6 shows the distribution of important features for the pooled models across sites. Both AKI 24h and 48h settings' pooled models have relatively fewer features that are only important at a small number of sites compared to the local model framework. For the pooled MLP and LR AKI 24h or 48h model, the universally important features mainly included *creatinine_last*, *potassium_max*, and *creatinine_min*. All AKI pooled models have features that are uniquely important to the pooled models (i.e., these features were not considered as part of the top 100 features at any local site). The pooled MLP AKI 24h model uniquely considered administration of Nitroglycerin moderately important. The pooled LR AKI 24h model uniquely considered administration of Metoclopramide Mupirocin, Lidocaine, and *race_black* as slightly important. The pooled MLP AKI 48h model uniquely considered administration of Hydromorphone and *bilirubin_last* as slightly and moderately important respectively. The pooled LR AKI 48h model also uniquely considered administration of Hydromorphone and

bilirubin_last as moderately important. Taken together, these differences suggested that there is slight variability in uniquely important features among models.

Fig 7 shows the distribution of important features for the federated models across sites. Like the pooled models, both MLP and LR federated models have relatively fewer features that are important at only a small number of sites compared to the local model analysis. The federated MLP AKI 24h model shares its universally important features with its pooled counterpart, namely attributing importance to *creatinine_last*, *potassium_max*, and *creatinine_min*, among others. These features are universally important in the federated LR architecture, as well as in the 48h setting. Some of the federated AKI models have uniquely important features as well. The federated MLP AKI 24h model considered administration of Dexmedetomidine as an important variable. The federated LR AKI 48h model considered administration of Phenylephrine slightly important. Like the pooled setting, we can see discrepancies between feature importances which are not considered universally important at all sites.

Correlation of feature importances across model architectures

To investigate differences of feature importance between model architectures, we looked at the correlation between importance rankings of features shared by both the MLP and LR model for each setting and framework. Fig 8 shows these correlations, where the x and y axes are the importance of the feature in the MLP and LR model respectively. In the AKI 24h setting, local models have moderately strong positive correlations with Pearson-correlation coefficients (PC) ranging from 0.79–0.84. The pooled and federated AKI 24h model shows slightly weaker positive correlations as compared to the local models PC = 0.79, 0.77. These results suggest that, within the AKI setting, the pooled and federated models were not successful at decreasing feature discrepancies between the LR and MLP architectures that were present in local models. Sepsis prediction setting results are available in Supplemental Information and S9 Fig.

Correlation between local feature importances and non-local framework feature importances

To investigate the correlation of heterogeneous features between local frameworks and both pooled and federated frameworks, we established the 'unique importance score' (UIS). The UIS score is large for features that are highly important at a small subset of sites, whereas it is small for features that are considered universally important (i.e., features that were important at a plurality of sites). In other words, the score is large for features that lie in the bottom right region of the plots in Figs 5, 6, and 7. Calculation of the UIS score can be found in the Methods section. Fig 9 shows the correlations of the UIS score across frameworks. Similar conclusions can be derived from both pooled and federated framework analyses, in both sepsis and AKI. There is a strong positive correlation (PC ranging from 0.84–0.93) between local UIS and pooled/federated. Interestingly, for all analyses, confidence on the line of best fit decreases at larger UIS scores. This suggests that features considered universally important in the local framework were important for the pooled/federated models whereas features only important at a small subset of hospitals were disregarded.

Sources of data heterogeneity

Tables 1 and 2 show demographic profiles across each site for AKI and sepsis patients, respectively. For both AKI and sepsis settings, sites show similar gender distributions, with a slight majority of patients being male across all sites. Age distributions are also similar across all sites with most patients being between 50–75 years old. Patient BMI is similar across sites with most patients having a BMI between 23–34. Site 199 has slightly fewer patients with a BMI of

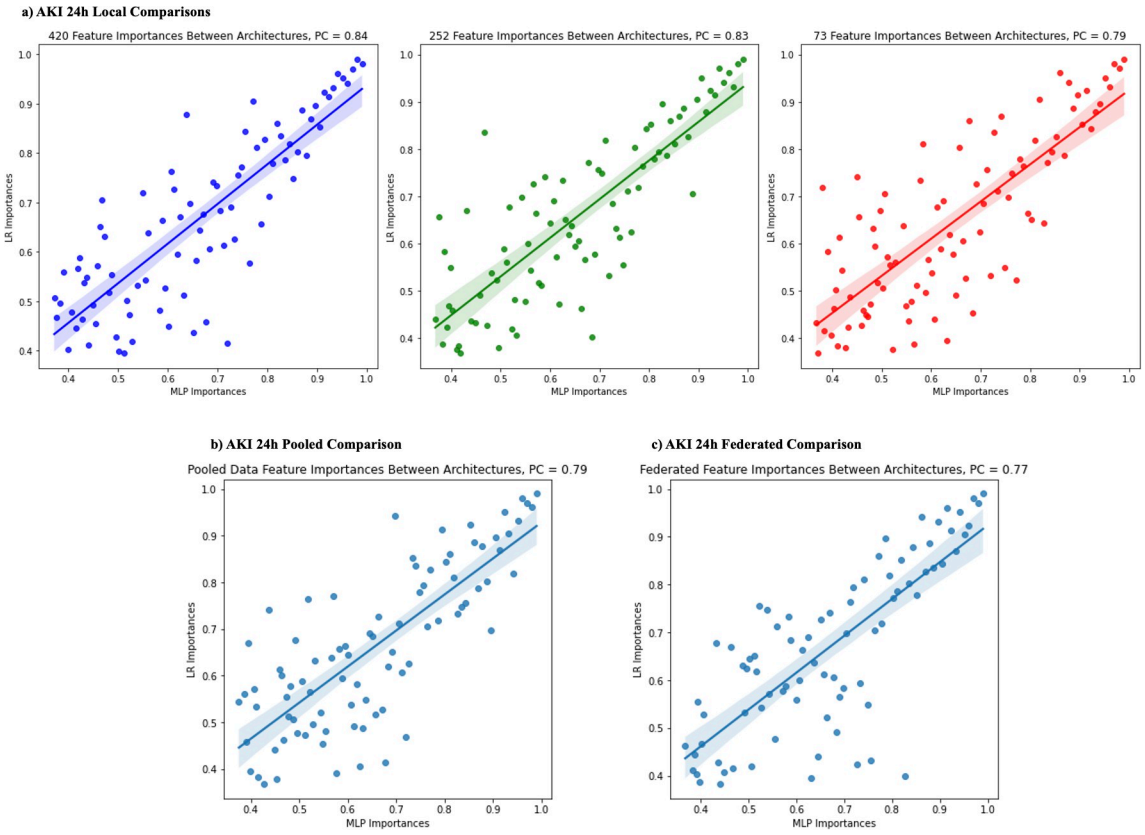

Fig 8. Important features comparison across model architectures for AKI 24h setting. The figure shows correlations between important features in the MLP and LR models. Each dot corresponds to one of the most important features ranked among the top-100 by both the MLP and LR model. The y-axis measures the importance of the feature in the LR model whereas the x-axis measures the importance in the MLP model. The shaded portion represents a 95% confidence interval. PC (Pearson correlation coefficient) for each comparison is denoted on the top-left of each plot. (a) shows the comparisons for local sites, 420, 252, and 73. (b) shows comparisons for the pooled models. (c) shows comparisons for the federated models.

https://doi.org/10.1371/journal.pdig.0000117.g008

less than 23, and more patients with a BMI of greater than 34 compared to other sites. In both settings, there was a disparity in the number of patients that underwent elective surgery, with the proportions ranging from 0.12–0.28. Patients show differences in racial breakdowns across sites. The African American population varies across sites from 0.02/0.01 (AKI/sepsis) at Site 199 to 0.3/0.32 at Site 243. Site 73 has a relatively large population of Hispanic individuals compared to other sites, whereas Sites 122, 243, 252, and 458 have no Hispanic patients. The Asian population is similar across all sites. The 'Other' racial category has the largest proportion of individuals across all sites, but this proportion varies largely depending on the site, ranging from 0.67–0.98. As previously mentioned, most of the patients in all settings (AKI 24h, 28h, and sepsis) were negative for the disease. For AKI 24h and AKI 48h settings, the proportion of AKI positive patients ranges from 0.06/0.08 (24h/48h) to 0.1/0.13. For the sepsis setting, the proportion of positive patients ranges from 0.02 to 0.20.

Table 3 shows general site information for the 7 hospitals. The sites are located across the Northeast, Midwest, and South of the continental United States of America. All sites are large with greater than 500 beds. There are differences in patient unit types across all sites. Sites 420, 243, 252, and 199 have no patients in Cardiothoracic Intensive Care Units (CTICU). Sites 252 and 458 have no patients in Medical Surgery Units (Med-Surg ICUs). Sites 122 and 199 have no patients in Surgical Intensive Care Units (SICU). Sites 122, 243, 458, and 199 have no

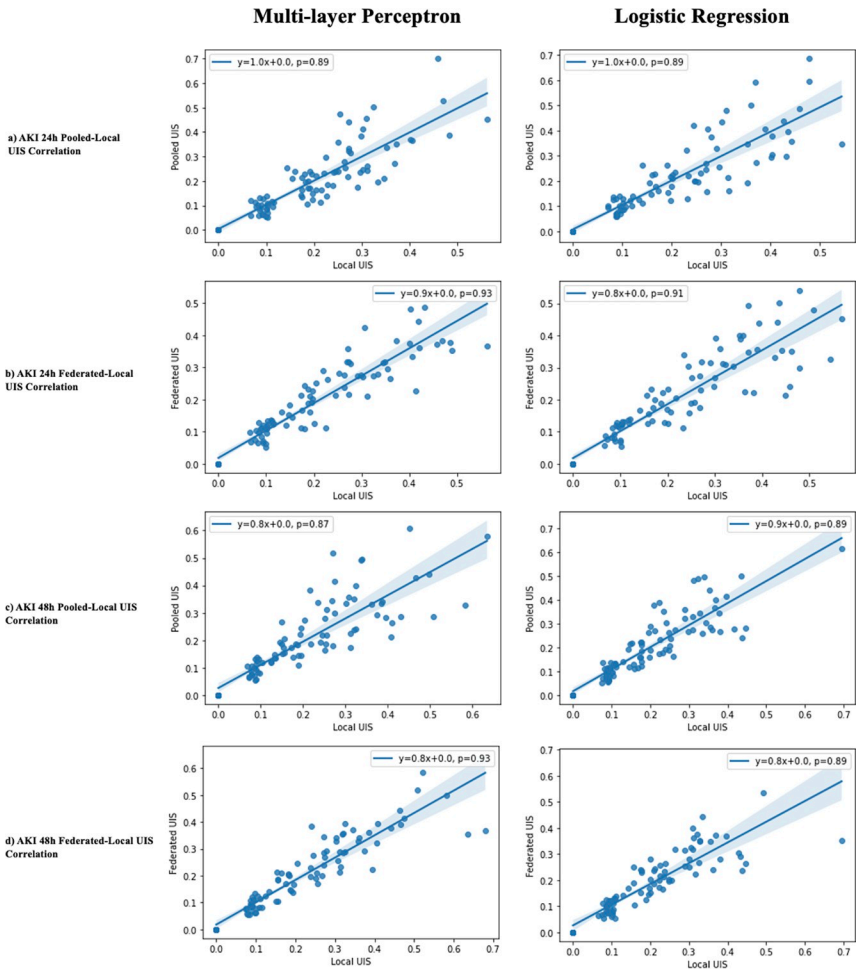

Fig 9. Correlation of unique importance score (UIS) between local and pooled/federated frameworks. x-axis is the UIS for each feature in the local model framework. y-axis is the UIS for the pooled/federated model. Line of best fit is plotted, and equation is shown on top corner, along with Pearson correlation coefficient (p). Shaded area represents a 95% confidence interval. First column depicts plots for the MLP model whereas the second column depicts plots for the LR model. (a, b) show analyses for sepsis setting, (c, d) show analyses for AKI 24h setting. (a, c) show analyses for pooled framework, (b, d) show analyses for federated framework.

https://doi.org/10.1371/journal.pdig.0000117.g009

patients in Critical Care Cardiothoracic Intensive Units (CCU-CTICU). Sites 420 and 122 have no patients in Cardiothoracic Intensive Care Units (MICU). Sites 420, 122, and 199 have no patients in Neurological Care Units (Neuro ICU). Sites 252, 199, and 73 have no patients in Cardiac Intensive Care Units (Cardiac ICU). Among sites which do share patients in the same unit, proportions may be different. For example, while both Sites 199 and 73 have patients in the Med-Surg ICUs, 88% of patients in Site 199 are admitted to Med-Surg ICUs whereas only 16% of patients in Site 73 are admitted to Med-Surg ICUs. Sites also had disparities in patient admission sources. Across hospitals, most patients were either admitted directly or admitted from the emergency department or operating Room. Of note, at Site 199, 22% of patients were admitted from the ICU to a special care unit (SCU). At Site 73, no patients were admitted from the recovery room. Taken together, despite being large, sites have disparities in unit types and sources of admission.

S5 Fig illustrates the usage of medications across sites. Only 22 medications are used at all sites for both sepsis and AKI settings. Further analysis indicated that even for medications

Table 1. Demographic Characteristics of AKI patients at each site. Percentage of individuals with certain characteristics specified within parentheses. Table 1 Positive/Negative distribution is associated with the AKI 48h setting.

	ID 420	ID 122	ID 243	ID 252	ID 458	ID 199	ID 73
Total	2957	2315	2990	2549	2592	2848	5545
Female	1198 (0.41)	1080 (0.47)	1316 (0.44)	1079 (0.42)	1161 (0.45)	1232 (0.43)	2452 (0.44)
Male	1759 (0.59)	1235 (0.53)	1674 (0.56)	1470 (0.58)	1431 (0.55)	1616 (0.57)	3093 (0.56)
Black	131 (0.04)	507 (0.22)	909 (0.3)	164 (0.06)	730 (0.28)	43 (0.02)	813 (0.15)
Hispanic	2 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	26 (0.01)	397 (0.07)
Asian	46 (0.02)	11 (0.0)	28 (0.01)	8 (0.0)	36 (0.01)	2 (0.0)	72 (0.01)
Other	2778 (0.94)	1797 (0.78)	2053 (0.69)	2377 (0.93)	1826 (0.7)	2777 (0.98)	4263 (0.77)
Elective Surgery: Yes	366 (0.12)	404 (0.17)	470 (0.16)	625 (0.25)	490 (0.19)	368 (0.13)	1542 (0.28)
Elective Surgery: No	2591 (0.88)	1911 (0.83)	2520 (0.84)	1924 (0.75)	2102 (0.81)	2480 (0.87)	4003 (0.72)
BMI < 23	582 (0.2)	493 (0.22)	493 (0.17)	480 (0.19)	553 (0.21)	399 (0.15)	806 (0.2)
23 < BMI < 28	959 (0.33)	678 (0.3)	884 (0.3)	743 (0.29)	792 (0.31)	785 (0.29)	1100 (0.27)
28 < BMI < 34	748 (0.26)	606 (0.27)	904 (0.3)	703 (0.28)	722 (0.28)	764 (0.28)	1121 (0.28)
BMI > 34	613 (0.21)	485 (0.21)	701 (0.24)	617 (0.24)	520 (0.2)	779 (0.29)	1010 (0.25)
Age < 25	107 (0.04)	71 (0.03)	73 (0.02)	81 (0.03)	97 (0.04)	118 (0.04)	116 (0.02)
25 < Age < 50	559 (0.19)	420 (0.18)	505 (0.17)	437 (0.17)	563 (0.22)	537 (0.19)	1032 (0.19)
50 < Age < 75	1611 (0.54)	1322 (0.57)	1757 (0.59)	1516 (0.59)	1414 (0.55)	1599 (0.56)	3096 (0.56)
Age > 75	680 (0.23)	502 (0.22)	655 (0.22)	515 (0.2)	518 (0.2)	594 (0.21)	1301 (0.23)
Positive	338 (0.11)	196 (0.08)	382 (0.13)	216 (0.08)	271 (0.1)	233 (0.08)	476 (0.09)
Negative	2619 (0.89)	2119 (0.92)	2608 (0.87)	2333 (0.92)	2321 (0.9)	2615 (0.92)	5069 (0.91)

https://doi.org/10.1371/journal.pdig.0000117.t001

used at multiple hospitals, the proportion of patients that were on the medication at each hospital varied greatly. Coupled with the disparities in unit types, this suggests that each hospital site treats significantly different populations of individuals, despite all these hospitals having patients who suffer from AKI and sepsis.

Table 2. Demographic Characteristics of Sepsis patients at each site. Percentage of individuals with certain characteristics specified within parentheses.

	ID 420	ID 122	ID 243	ID 252	ID 458	ID 199	ID 73
Total	2276	2365	3212	2586	2748	2996	5919
Female	936 (0.41)	1085 (0.46)	1404 (0.44)	1097 (0.42)	1234 (0.45)	1293 (0.43)	2629 (0.44)
Male	1340 (0.59)	1280 (0.54)	1808 (0.56)	1489 (0.58)	1514 (0.55)	1703 (0.57)	3290 (0.56)
Black	100 (0.04)	553 (0.23)	1019 (0.32)	175 (0.07)	805 (0.29)	44 (0.01)	899 (0.15)
Hispanic	2 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	36 (0.01)	428 (0.07)
Asian	43 (0.02)	12 (0.01)	31 (0.01)	10 (0.0)	38 (0.01)	3 (0.0)	82 (0.01)
Other	2131 (0.94)	1800 (0.76)	2162 (0.67)	2401 (0.93)	1905 (0.69)	2913 (0.97)	4510 (0.76)
Elective Surgery: Yes	371 (0.16)	424 (0.18)	541 (0.17)	659 (0.25)	527 (0.19)	412 (0.14)	1635 (0.28)
Elective Surgery: No	1905 (0.84)	1941 (0.82)	2671 (0.83)	1927 (0.75)	2221 (0.81)	2584 (0.86)	4284 (0.72)
BMI < 23	414 (0.19)	491 (0.21)	528 (0.16)	475 (0.18)	576 (0.21)	414 (0.14)	854 (0.2)
23 < BMI < 28	747 (0.33)	692 (0.3)	953 (0.3)	761 (0.29)	809 (0.3)	801 (0.28)	1170 (0.27)
28 < BMI < 34	597 (0.27)	628 (0.27)	956 (0.3)	723 (0.28)	775 (0.28)	811 (0.28)	1188 (0.28)
BMI > 34	475 (0.21)	496 (0.21)	765 (0.24)	621 (0.24)	582 (0.21)	848 (0.3)	1100 (0.26)
Age < 25	77 (0.03)	72 (0.03)	72 (0.02)	83 (0.03)	100 (0.04)	119 (0.04)	125 (0.02)
25 < Age < 50	430 (0.19)	435 (0.18)	542 (0.17)	448 (0.17)	577 (0.21)	557 (0.19)	1083 (0.18)
50 < Age < 75	1243 (0.55)	1358 (0.57)	1902 (0.59)	1516 (0.59)	1516 (0.55)	1694 (0.57)	3318 (0.56)
Age > 75	526 (0.23)	500 (0.21)	696 (0.22)	539 (0.21)	555 (0.2)	626 (0.21)	1393 (0.24)
Positive	459 (0.2)	172 (0.07)	187 (0.06)	152 (0.06)	123 (0.04)	135 (0.05)	89 (0.02)
Negative	1817 (0.8)	2193 (0.93)	3025 (0.94)	2434 (0.94)	2625 (0.96)	2861 (0.95)	5830 (0.98)

https://doi.org/10.1371/journal.pdig.0000117.t002

Table 3. Hospital site information. Percentage of usage/individuals specified within parentheses.

	ID 420	ID 122	ID 243	ID 252	ID 458	ID 199	ID 73
Region	Northeast	South	South	Midwest	South	Northeast	Midwest
Number of Beds	> = 500	> = 500	> = 500	> = 500	> = 500	> = 500	> = 500
Unit Type							
CTICU	0 (0.0)	562 (0.19)	0 (0.0)	0 (0.0)	635 (0.17)	0 (0.0)	1364 (0.19)
Med-Surg ICU	1343 (0.29)	1495 (0.51)	11 (0.0)	0 (0.0)	0 (0.0)	3712 (0.88)	1159 (0.16)
SICU	1932 (0.41)	0 (0.0)	710 (0.17)	588 (0.17)	372 (0.1)	0 (0.0)	408 (0.06)
CCU-CTICU	706 (0.15)	0 (0.0)	0 (0.0)	1031 (0.31)	0 (0.0)	0 (0.0)	1510 (0.21)
MICU	0 (0.0)	0 (0.0)	778 (0.18)	914 (0.27)	446 (0.12)	528 (0.12)	1124 (0.16)
Neuro ICU	0 (0.0)	0 (0.0)	716 (0.17)	838 (0.25)	416 (0.11)	0 (0.0)	1494 (0.21)
Cardiac ICU	698 (0.15)	884 (0.3)	2028 (0.48)	0 (0.0)	1832 (0.5)	0 (0.0)	0 (0.0)
Patient Admit Source							
Floor	683 (0.15)	676 (0.23)	378 (0.09)	175 (0.05)	261 (0.07)	290 (0.07)	904 (0.13)
Emergency Department	2283 (0.5)	1254 (0.43)	1798 (0.42)	760 (0.23)	1463 (0.4)	1185 (0.28)	2301 (0.33)
Operating Room	644 (0.14)	458 (0.16)	402 (0.09)	462 (0.14)	577 (0.16)	698 (0.17)	1858 (0.26)
Direct Admit	49 (0.01)	397 (0.14)	763 (0.18)	877 (0.26)	710 (0.19)	427 (0.1)	627 (0.09)
Other Hospital	279 (0.06)	27 (0.01)	19 (0.0)	100 (0.03)	82 (0.02)	96 (0.02)	233 (0.03)
ICU to SDU	178 (0.04)	26 (0.01)	234 (0.06)	0 (0.0)	1 (0.0)	916 (0.22)	870 (0.12)
Other ICU	156 (0.03)	67 (0.02)	78 (0.02)	267 (0.08)	124 (0.03)	138 (0.03)	194 (0.03)
Step-Down Unit (SDU)	94 (0.02)	3 (0.0)	189 (0.04)	417 (0.12)	251 (0.07)	432 (0.1)	49 (0.01)
Chest Pain Center	103 (0.02)	5 (0.0)	0 (0.0)	0 (0.0)	11 (0.0)	0 (0.0)	1 (0.0)
Recovery Room	140 (0.03)	20 (0.01)	372 (0.09)	310 (0.09)	209 (0.06)	21 (0.0)	0 (0.0)
Acute Care/Floor	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	11 (0.0)	0 (0.0)
PACU	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	3 (0.0)	0 (0.0)
ICU	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	10 (0.0)	0 (0.0)

https://doi.org/10.1371/journal.pdig.0000117.t003

Discussion

In this study, to investigate the effect of data heterogeneity on the performance of FL, multiple machine learning models were developed to predict the risk of both AKI and sepsis diseases in multiple ICU settings. Different types of EHRs including lab tests, vital signs, demographics, and medications were extracted from seven hospitals in the eICU collaborative research database. Three model frameworks including local, pooled, and federated were explored. Effects of data heterogeneity across hospital sites were evaluated through model performance comparison and feature importance analysis. The sources of data heterogeneity across hospitals were investigated based on patient demographics, medication usage, and general hospital attributes.

Our prediction models have shown comparable performance with state-of-the art AKI and sepsis prediction studies [7]. In addition, federated model frameworks generally outperform their local counterparts in our results. However, this largely depends on how heterogeneous the patient populations from different hospitals are. Moreover, the pooled model did not show much improvement over the local models, this could be largely due to the cross-site sample heterogeneity. Though FL performed better than pooled models in our investigations, our FL strategy is based on federated average which did not consider such cross-site heterogeneities, thus it is difficult to justify the generalizability of the conclusion. One reason as to why the federated models performed better than pooled models might be due to their weight distribution. The weights of the federated models are concentrated around zero as compared to pooled models for all settings. More weights near zero means that the models are regularized and simpler, which is likely to generalize better [29].

The performance heterogeneity of predictive models across sites and frameworks was evaluated by comparing feature importance. For both AKI and sepsis prediction tasks, important variables identified by predictive models were consistent with prior studies [7]. For example, creatinine and furosemide exposure showed positive associations with AKI, which is unsurprising given their clinical association with AKI. For the same model architecture, importance of a feature varied depending on the sites, with variable-prediction relationships changing (see Fig 5, 6, 7). The presence of 'universally important features' (i.e., features that were considered highly important at most sites) and 'uniquely important features' (i.e., features that were highly important at a small subset of sites) showed that there was disagreement on relative importance across sites. The feature heterogeneity plots for federated and pooled frameworks showed a decreased amount of uniquely important features. This was indicative of both these frameworks being able to attribute higher importance to features shared across multiple sites.

Our findings also demonstrated that federated and pooled models were not successful at decreasing feature importance discrepancies between LR and MLP architectures, and that both pooled and federated frameworks prioritized features that were considered important across a plurality of sites (i.e., low UIS) and attributed lower importance to features that were uniquely important at a small subset of sites (i.e., high UIS). These findings also suggest that the federated model may be better at discriminating the key features of patient-level clinical, lab, and demographic information that improves risk prediction. In the field of critical care medicine, the implication of this finding is that across heterogeneous sources of data, federated models are more likely to highlight the common elements that can better predict sepsis and AKI between hospital, patient, and practice-specific circumstances, thus highlighting the generalizability of the model's value. However, consequently, it is possible that important local characteristics that may better predict AKI and sepsis within hospitals could be overlooked when compared to pooled or local models, which may in turn limit the clinical utility of these tools, a finding that is increasingly being acknowledged in the AI/ML literature.

Within our analysis, differences in features of the hospitals, and ICUs were notable. Many sites did not have any patients admitted to ICU types that other sites had a high proportion of patients within, for example the Medical Surgery ICU or SICU. At these hospitals, different sites treat different conditions. Thus, treatments may vary depending on the etiology and nature of the condition driving sepsis and AKI [30]. For example, a patient managed for decompensated heart failure in a cardiac ICU who subsequently develops AKI may be treated with inotropic support and furosemide, whereas a patient being managed for septic shock in a medical ICU with AKI may be aggressively repleted with intravenous fluids. As such different hospitals, which are specialized at treating different conditions, may have slightly differing medication regimens for treating patients when faced with the same disease, which in turn may be a function of the practicing physician and their choice of treatment options including medications, prespecified protocols, or even higher-level decisions about cost within centralized hospital pharmacies. Our models highlighted this putative disagreement between medication usage at hospitals, creating another source of heterogeneity in model training. This heterogeneity in medications and demographic details was demonstrated in the feature heterogeneity plots since features with higher UIS scores tend to be medications and demographic information. Differently, lab tests and vital signs were generally universally important features across hospitals, likely because these are commonly standardized across hospitals. Taken together, local frameworks may heavily suffer in generalizability even when population demographics are similar across sites, due to disagreements in medication and treatment administration. However, clinicians may find use in site-specific factors, which may not be evident in federated frameworks and only be ascertainable within a local framework. As such, while

federated frameworks may provide performance increases, local frameworks can still provide clinical value in determining important site-specific factors for risk prediction.

Limitations

There are several limitations to our study. First, we mainly considered structured clinical information to construct the features. Integrating unstructured free text to build predictive models may obtain better predictive performance and allow a new level of explainability of prediction. Second, we only considered LR and MLP to build predictive models based on local, pooled, and federated frameworks. Other algorithmic solutions such as support vector machines may have a potential to improve model performance. Third, we mainly focus on describing the effects of data heterogeneity in FL in terms of disease risk prediction. Considering data harmonization techniques and other federated techniques to mitigate the problem and improve the performance is one of future research topics. Moreover, federated techniques that deal with data heterogeneity while simultaneously reducing communication costs may be required for real-time medical use.

Methods

Ethics statement

This study analyzed a publicly available anonymized database (eICU Collaborative Research Database) with preexisting institutional review board approval. Collection of data was in accordance with the ethical standards set out by the IRB no. 0403000206 of the Massachusetts Institute of Technology and with the 1964 Declaration of Helsinki and its later amendments. Because the database is fully anonymized, formal consent was not required to use the data.

Data aggregation

Patient data was extracted from the eICU Collaborative Research Database, a multi-center critical care database made publicly available through Philips Healthcare and the MIT Laboratory for Computational Physiology (https://eicu-crd.mit.edu/). The database contains detailed information regarding the clinical care of ICU patients. We investigated three disease settings (24h or 48h observation window (OW) AKI, and sepsis). An AKI (non-graded) is defined as any of the following: Increase in serum creatinine (SCr) by $>=$ 0.3 mg/dl ($>=$ 26.5 μmol/l) within 48 hours, increase in SCr to $>=$ 1.5 times baseline which is known or presumed to have occurred within the prior 7 days, or urine volume $<$ 0.5 ml/kg/h for 6 hours. We predict AKI risk using an accumulating OW (S1 Fig). We predicted AKI within the next 24 hours (prediction window, PW) following the end of the OW, focusing only on the first 3 days (72 hours) of a patients' inpatient hospital stay (max OW = 48 hours). For each patient, we created 2 pairs of OW/PWs, specifically using from OW = 1–24 hours (1-day) after admission, 1–48 hours (2-days). We do not consider the onset point. For the AKI prediction experimental setting, positive cases are samples that are diagnosed as AKI in the prediction window whereas controls are samples that are not diagnosed as AKI in the prediction window. For sepsis prediction, we labeled patient data in accordance with Sepsis-3 clinical criteria. We predicted whether patients would suffer from sepsis 6 hours prior to onset, onset point included. For sepsis prediction experimental setting, positive cases are samples that are diagnosed as sepsis. Controls are samples that are not diagnosed as sepsis. For patients who did not develop sepsis, predictor values were selected from a random T-hour time window (T is usually set as 48 or 24 hours) during the patient's ICU stay. For those who developed sepsis, a time was selected for the patient within admission to 6 hours prior to the onset of sepsis, and the predictor values

were extracted. Data was collected from seven hospitals with the following IDs: 420, 122, 243, 252, 458, 199, and 73. For all three disease predictions (24h or 48h AKI, and sepsis), all hospital sites shared all features including: general demographic information (8 variables), vital signs/lab tests (29 variables), and medications (254 medications). For 28 vital signs and lab tests, the max, min, first, and last values are calculated. For urine, only the summation is calculated. Taken together, a total of 354 features were available at every hospital site for each patient.

Data processing

For all datasets, we performed an automated curation process outlined as follows: (1) systematically identified extreme values of numerical features (e.g., vital signs/lab tests and some demographic information) that were beyond the 1st and 99th percentile as outliers. We marked these values as missing. Primarily, this step marked values within demographic data (BMI, age) and some vital signs as missing. Values marked as missing were investigated through clinical literature to confirm that they were physiologically impossible. Previous studies utilizing the eICU Collaborative Research Database have noted these errors are at random and can be removed in downstream analyses [31–32]. (2) We standardized all our variables appropriately by normalizing all our numerical features and converting binary features to either 1 or -1. (3) For all missing measurements, the Multiple Imputation by Chained Equations algorithm (MICE) was used. MICE imputation can calculate missing information by taking advantage of the relationships between non-missing measurements within the dataset. Because overall patient distributions are conserved after outlier removal (due to limited number of values being considered outliers), MICE imputation can provide robust estimation of these values as well [33].

Experimental design

There were three prediction tasks including 24-hour and 48-hour prediction of AKI, and sepsis prediction. Three model frameworks were designed including local, pooled, and federated model frameworks. The local model framework only used data from each site itself. The pooled model framework combined data from all sites. In the federated model framework, each local site does not have access to other sites' data. A model was trained locally, and its parameters were shared to a central aggregator, which was used to update global model parameters which were subsequently sent back to each site. For each framework, LR and MLP were used as model architectures, so there are 54 tasks in total were performed (7 site-specific (local) x 3 prediction tasks x 2 architectures + 1 pooled model x 3 prediction tasks x 2 architectures + 1 federated model x 3 prediction tasks x 2 architectures). For all settings, five-fold cross-validation was used during training models. The Shapely Additive exPlanations (SHAP) tool was used to calculate feature importance rankings for each task. The Markov Chain Type 4 rank aggregation was used to combine the feature importance rankings for all five folds.

Learning algorithm

To investigate the effects of heterogeneity across architectures, we focused on two learning models: multilayer perceptron (MLP) and logistic regression (LR). The MLP is a class of feedforward artificial neural network (ANN) with a non-parametric functional form [34]. An MLP consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable. Since MLPs are fully connected, each node in one layer

connects with a certain weight to every node in the following layer [35]. To implement the MLP model, Python's PyTorch library was used. PyTorch is an open-source machine learning framework based on the Torch library, used for applications such as computer vision and natural language processing, primarily developed by Facebook's AI Research lab [36]. All MLP models had one hidden dense layer of 10 units, learning rate = 0.001, used binary cross-entropy loss, and stochastic gradient descent optimization. To mitigate class imbalance, class weights were used to penalize the loss for positive class inaccuracies. This allows the model to pay increased attention to examples from the positive class despite a skewed class distribution [37]. Each model was trained for 200 epochs and the batch size was 64. An epoch is the total number of iterations it takes all the training data to make one pass through the model whereas, the batch size is the number of samples processed in each iteration before the model is updated [38].

In addition to the MLP model, we implemented a LR model. The LR model has a parametric functional form and formulates the log-odds of an event as a linear combination of independent variables [39]. The LR model consists of one linear layer followed by sigmoid activation. Like the MLP, a learning rate = 0.001, binary cross-entropy loss, and stochastic gradient descent optimization was used. Class weights were applied in a similar fashion to the MLP model. For consistency and to enable direct comparisons, all models of each framework for all tasks were built with the same architecture.

Because the output of an MLP model is a nonlinear function of the inputs, the decision boundary for classification from an MLP is also nonlinear, which provides more flexibility than LR models [34]. As such, we wanted to investigate the effects of heterogeneity across these two different architectures.

Our primary model framework of interest was a federated learning model. In this model, training was performed in different sites, and parameters were shared to a central location. To create a federated model using both the MLP and LR architecture, the federated averaging technique was used. The process was as follows: a central aggregator initialized the federated model with random parameters. This model was sent to each site, then trained for one epoch. Next, model parameters were sent back to the central aggregator where federated averaging was performed. Updated parameters from the central aggregator were then sent back to each site, and this cycle was repeated for multiple epochs. Federated averaging scales the parameters of each site according to the number of available data points and sums all parameters by layer. Through this technique, federated models did not receive any raw data. Class weighting was performed at each site on every cycle, which ensured local data distribution information was not sent to the global server. All parameters for local server models were kept the same to enable comparison. We were able to perform federated class weighting through this mechanism because local data distributions were similar across hospitals.

Assume M local sites, each with N_m samples (number of samples in m-th local site). $w_n^{(m)}$ is the weight of the n-th sample at the m-th site, $y_n^{(m)}$ is the ground-truth label for sample $x_n^{(m)}$, which is the n-th sample at m-th site:

$$\{x_n^{(m)}, y_n^{(m)} w_n^{(m)}\}_{m=1,n=1}^{M,N_m}$$

The base model Ω is initialized at the global server. Without loss of generalizability, the following steps assume a LR model described by Eq (1). $\widehat{y_n}$ is the predicted value for sample x_n by the LR model with parameters β:

$$\widehat{y_n} = \frac{1}{1 + e^{-\beta x_n}} \tag{1}$$

At the start of each epoch, model Ω is copied from the global server to the local site Ω_m:

$$\Omega_m \leftarrow \Omega$$

Ω_m is trained on local site data for one epoch. Loss at the m-th site (l_m) is calculated using modified binary-cross entropy which considers class weighting at site, described by Eq (2). $\widehat{y_n^{(m)}}$ is the model prediction for sample $x_n^{(m)}$ and *bce()* is the binary cross-entropy function:

$$l_m = \sum_{n=1}^{N_m} w_n^{(m)} bce(y_n^{(m)}, \widehat{y_n^{(m)}}) \tag{2}$$

Local model Ω_m is updated through back propagation and gradient descent:

$$\Omega_m \leftarrow \Omega_m - \frac{\partial l_m}{\partial \Omega_m}$$

All β_m are transferred to the global server where the layer weights of all β_m are averaged through Eq (3), generating an updated global model for the next epoch:

$$\Omega = \frac{1}{M} \sum_{m=1}^{M} \Omega_m \tag{3}$$

The time complexity of one iteration of federated averaging is $O(Z_m N_m)$ for client m, where Z_m is the number of parameters in the model. The communication cost of one iteration is $O(Z_m)$.

Evaluations

We used the area under the receiver operator curve (AUROC) to compare the overall prediction performance, which is known to be more robust to imbalanced datasets. In addition to AUROC, accuracy, precision, and recall were calculated. In addition to aggregate performance metrics for each model, training loss and training/testing AUROC histories were measured. Tests of significance were performed using the student's t-test. Feature importance rankings for each task were computed using SHAP. To focus more on the most impactful features (i.e., variables ranked among top 100) without losing information on the weaker features, we assigned a "soft" membership of a feature as how high up the rank is relative to tops (s = 100) by applying an exponentially decreasing function to the original ranks (r), i.e., $f(r) = \exp\{-r/s\}$. For some top features, SHAP dependence plots were generated to illustrate the effect that each feature has on the predictions made by the model. Locally Weighted Scatterplot Smoothing (LOWESS) was used to fit a smooth trend line to the dependence plots.

The unique importance score (UIS) was calculated for each model architecture-setting-framework combination. For local model analysis, the mean importance i_{lj} for each feature j was calculated by averaging all soft rankings for said feature across all sites. This was done for all top 100 features at each local site. For both pooled and federated analysis, importance (i_{pj} or i_{fj}) of each feature j was set to the soft ranking of said feature within the pooled or federated model. In all model frameworks, the frequency f of each feature was calculated by determining the proportion of local sites the feature was a top 100 feature. Given i_{lj}, i_{pj}, i_{fj}, and f:

$$\mathrm{UIS}_{\mathrm{local,j}} = i_{lj} \text{ x } (1 - f_j)$$

$$\mathrm{UIS}_{\mathrm{pooled,j}} = i_{pj} \text{ x } (1 - f_j)$$

$$\mathrm{UIS}_{\mathrm{federated,j}} = i_{fj} \text{ x } (1 - f_j)$$

Supporting information

S1 Table. Performance summaries of all sepsis models. Performances of LR and MLP models are shown for each model framework. Four metrics are captured: accuracy, AUC, precision, and recall.
(DOCX)

S2 Table. Performance summaries of all AKI 24h models. Performances of LR and MLP models are shown for each model framework. Four metrics are captured: accuracy, AUC, precision, and recall.
(DOCX)

S3 Table. Performance summaries of all AKI 48h models. Performances of LR and MLP models are shown for each model framework. Four metrics are captured: accuracy, AUC, precision, and recall.
(DOCX)

S1 Fig. Prediction setting details of AKI and Sepsis. For AKI prediction, there are two observation windows (2 OWs) which creates 2 AKI prediction settings. The observation window for AKI settings can be 24 or 48 hours. For sepsis prediction, the observation window is the entire period from admission to 6 hours prior to the onset of sepsis.
(TIF)

S2 Fig. Sepsis training and testing histories for local and federated models. Training AUC, testing AUC, and training loss at each epoch (from left to right) has been shown. (a, c) show local histories where each color indicates the histories of a different site. (b, d) show histories for the federated model, where each color shows the history of the model while training/testing on that site's data. Training and testing histories for AKI settings show similar patterns to the sepsis setting.
(TIF)

S3 Fig. AKI 24h training and testing histories for local and federated models. Training AUC, testing AUC, and training loss at each epoch (from left to right) has been shown. (a, c) show local histories where each color indicates the histories of a different site. (b, d) show histories for the federated model, where each color shows the history of the model while training/testing on that site's data. Training and testing histories for AKI settings show similar patterns to the sepsis setting.
(TIF)

S4 Fig. Shapley dependence plots for 10 important features for local models. Each panel shows the marginal effects of impactful features for predicting sepsis, AKI 24h, or 48h in all local site models. All 7 sites are plotted in each panel, where each color corresponds to a different site (see legend). The x-axis gives the raw values of each feature, and the y-axis gives the logarithmic of estimated odds ratio (i.e., the SHAP value) for sepsis, AKI 24h or AKI 48h, when a feature takes a certain value. Each dot represents the SHAP value of a sample. The LOWESS curve, used for smoother extrapolating across all the dots, is plotted in all panels for each site. (a, c, and e) show Shapley dependence plots for federated MLP models and (b, d, and f) show Shapley dependence plots for federated LR models. (a, b) show plots for sepsis, (c, d)

show plots for AKI 24h, and (e, f) show plots for AKI 48h.
(TIF)

S5 Fig. Medication usage across local sites. (a, b) shows frequency of medications across hospitals. X-axis is the number of hospitals and y-axis is the number of medications. For example, there are ~20 medications that only appear at 1 hospital. (c, d) show disagreement of medication usage across hospitals for medications that appear at 2 or more hospitals. X-axis shows the standard deviation bins of proportions of patients using the medication at each hospital (i.e., larger values of standard deviation indicate more disagreement). Y-axis shows the number of medications within the histogram bin.
(TIF)

S6 Fig. Area Under Receiver Operating Curve (AUROC) for sepsis Setting. Each plot shows performances for the sepsis prediction setting. Blue bars depict each local site's model performance on their respective site test data. Orange bars depict pooled model performance on each local site's test data. Green bars depict federated model performance on each local site's test data.
(TIF)

S7 Fig. Shapley dependence plots for top 10 features for pooled and federated sepsis models. Each panel shows the marginal effects of each of the most impactful features ranked among the top 10 for predicting sepsis using pooled and federated models. The x-axis gives the raw values of each feature, and the y-axis gives the logarithmic of estimated odds ratio (i.e., the SHAP value) for sepsis when a feature takes a certain value. Each dot represents the SHAP value of a sample. The LOWESS curve, used for smoother extrapolating across all the dots, is plotted in red for all panels. (a, c) show Shapley dependence plots for MLP models and (b, d) show Shapley dependence plots for LR models. (a, b) show plots for pooled models, (c, d) show plots for federated models.
(TIF)

S8 Fig. Distribution of important features at all local, pooled, and federated models across local sites. The figure demonstrates feature importance disparities for the sepsis setting and model architectures (MLP and LR). (a-c) show feature importance disparities for MLP models. (d-f) show feature importances for LR models. Each dot corresponds to one of the most important features ranked among the top-100 by at least one of the seven models; y-axis measures the proportions of sites that identified the feature as top-100, or "commonality across sites"; x-axis measures the mean of feature importance rankings measured as "soft ranking" (the closer it is to 1, the higher the feature ranks). Top-100 is an arbitrary cutoff we used to analyze the most important features to illustrate heterogeneity. In (a, d) each feature is also color coded by the interquartile range (IQR) of the ranks across sites (the higher the IQR is, the more disagreement across sites on the importance of that feature). (b, e) show the most important features for the pooled models. (c, f) show the most important features for the federated models.
(TIF)

S9 Fig. Important features comparison across model architectures for sepsis setting. The figure shows correlations between important features in the MLP and LR models. Each dot corresponds to one of the most important features ranked among the top-100 by both the MLP and LR model. The y-axis measures the importance of the feature in the LR model whereas the x-axis measures the importance in the MLP model. The shaded portion represents a 95% confidence interval. PC (Pearson correlation coefficient) for each comparison is denoted

on the top-left of each plot. (a) shows the comparisons for local sites, 420, 252, and 73. (b) shows comparisons for the pooled models. (c) shows comparisons for the federated models.
(TIF)

S1 Text. Abbreviations.
(DOCX)

S2 Text. Validation on Sepsis Prediction Setting.
(DOCX)

Author Contributions

Conceptualization: Suraj Rajendran, Fei Wang.

Data curation: Zhenxing Xu.

Formal analysis: Suraj Rajendran.

Funding acquisition: Fei Wang.

Investigation: Suraj Rajendran, Zhenxing Xu, Weishen Pan.

Methodology: Suraj Rajendran.

Project administration: Fei Wang.

Software: Suraj Rajendran.

Supervision: Zhenxing Xu, Fei Wang.

Validation: Suraj Rajendran, Zhenxing Xu.

Visualization: Suraj Rajendran.

Writing – original draft: Suraj Rajendran, Zhenxing Xu.

Writing – review & editing: Suraj Rajendran, Zhenxing Xu, Weishen Pan, Arnab Ghosh, Fei Wang.

References

1. Zeng X., McMahon G. M., Brunelli S. M., Bates D. W., Waikar S. S. Incidence, outcomes, and comparisons across definitions of AKI in hospitalized individuals. https://doi.org/10.2215/CJN.02730313 PMID: 24178971
2. Rhee C, Dantes R, Epstein L, et al. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014. JAMA. 2017; 318(13):1241–1249. https://doi.org/10.1001/jama.2017.13836 PMID: 28903154
3. Cheng Peng, Waitman Lemuel R., Hu Yong, Liu Mei, Predicting inpatient acute kidney injury over different time horizons: How early and accurate? in: AMIA Annual Symposium Proceedings, vol. 2017, 2017, p. 565. PMID: 29854121
4. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA. 2016; 315(8): 762–774. https://doi.org/10.1001/jama.2016.0288 PMID: 26903335
5. Alobaidi R., Basu R. K., Goldstein S. L., Bagshaw S. M. Sepsis-associated acute kidney injury. Seminars in nephrology. 2015; 35(1), 2–11. https://doi.org/10.1016/j.semnephrol.2015.01.002 PMID: 25795495
6. Vincent J. L., Pereira A. J., Gleeson J., Backer D. Early management of sepsis. Clinical and experimental emergency medicine. 2014; 1(1), 3–7. https://doi.org/10.15441/ceem.14.005 PMID: 27752546
7. Song X, Yu AS, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nature Communications. 2020; 11(1). https://doi.org/10.1038/s41467-020-19551-w PMID: 33168827

8. Sarnowski A, Hodgson L. Systematic review of prognostic prediction models for acute kidney injury in general hospital populations: Methodology. 2020;
9. Koyner JL, Carey KA, Edelson DP, Churpek MM. The development of a machine learning inpatient acute kidney injury prediction model*. Critical Care Medicine. 2018; 46(7):1070–7. https://doi.org/10.1097/CCM.0000000000003123 PMID: 29596073
10. Churpek MM, Carey KA, Edelson DP, Singh T, Astor BC, Gilbert ER, et al. Internal and external validation of a machine learning risk score for Acute Kidney Injury. JAMA Network Open. 2020; 3(8). https://doi.org/10.1001/jamanetworkopen.2020.12892 PMID: 32780123
11. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine. 2021; https://doi.org/10.1001/jamainternmed.2021.2626 PMID: 34152373
12. Reyna M, Prajwal Shashikumar S, Moody B, Gu P, Sharma A, shamim nemati, et al. Early prediction of sepsis from clinical data: The PHYSIONET/computing in cardiology challenge 2019. 2019 Computing in Cardiology Conference (CinC). 2019;
13. Yan MY, Gustad LT, Nytrø Ø. Sepsis prediction, early detection, and identification using clinical text for Machine Learning: A Systematic Review. Journal of the American Medical Informatics Association. 2021; 29(3):559–75.
14. Yang Jenny, Andrew AS Soltan David A. Clifton. "Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening." npj Digital Medicine 5, no. 1 2022; 1–8.
15. Singh H, Mhasawade V, Chunara R. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. PLOS Digit Health. 2022; 1(4): e0000023. https://doi.org/10.1371/journal.pdig.0000023 PMID: 36812510
16. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports. 2020; 10 (1).
17. Xu Y, Ma L, Yang F, Chen Y, Ma K, Yang J, et al. A collaborative online AI engine for CT-based COVID-19 diagnosis. medRxiv Preprint posted online on May 19, 2020. https://doi.org/10.1101/2020.05.10.20096073 PMID: 32511484
18. Xu Jie, Glicksberg Benjamin S., Su Chang, Walker Peter, Bian Jiang, and Wang Fei. "Federated learning for healthcare informatics." Journal of Healthcare Informatics Research 5, no. 1. 2021; 1–19. https://doi.org/10.1007/s41666-020-00082-4 PMID: 33204939
19. Li Tian, Anit Kumar Sahu Ameet Talwalkar, Smith Virginia. "Federated learning: Challenges, methods, and future directions." IEEE Signal Processing Magazine 37, no. 3. 2020; 50–60.
20. Kuno T, Mikami T, Sahashi Y, Numasawa Y, Suzuki M, Noma S, et al. Machine learning prediction model of acute kidney injury after percutaneous coronary intervention. Scientific Reports. 2022; 12(1). https://doi.org/10.1038/s41598-021-04372-8 PMID: 35031637
21. Heung M, Koyner JL. Entanglement of sepsis, chronic kidney disease, and other comorbidities in patients who develop acute kidney injury. Seminars in Nephrology. 2015; 35(1):23–37. https://doi.org/10.1016/j.semnephrol.2015.01.004 PMID: 25795497
22. Vagliano I, Chesnaye NC, Leopold JH, Jager KJ, Abu-Hanna A, Schut MC. Machine learning models for predicting Acute kidney injury: A systematic review and critical appraisal. Clinical Kidney Journal. 2022; 15(12):2266–80. https://doi.org/10.1093/ckj/sfac181 PMID: 36381375
23. Tan AZ, Yu H, Cui L, Yang Q. Towards personalized federated learning. IEEE Transactions on Neural Networks and Learning Systems. 2022;1–17. https://doi.org/10.1109/TNNLS.2022.3160699 PMID: 35344498
24. Dinh CT, Vu TT, Tran NH, Dao MN, Zhang H. A new look and convergence rate of federated multitask learning with laplacian regularization. IEEE Transactions on Neural Networks and Learning Systems. 2022;1–11.
25. Xing H, Xiao Z, Qu R, Zhu Z, Zhao B. An efficient federated distillation learning system for Multitask Time Series classification. IEEE Transactions on Instrumentation and Measurement. 2022; 71:1–12.
26. Zhou P, Lin Q, Loghin D, Ooi BC, Wu Y, Yu H. Communication-efficient decentralized machine learning over heterogeneous networks. 2021 IEEE 37th International Conference on Data Engineering (ICDE). 2021;
27. Crowson MG, Moukheiber D, Arévalo AR, Lam BD, Mantena S, Rana A, et al. A systematic review of Federated Learning Applications for Biomedical Data. PLOS Digital Health. 2022; 1(5). https://doi.org/10.1371/journal.pdig.0000033 PMID: 36812504
28. Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, et al. Federated learning of Electronic Health Records to improve mortality prediction in hospitalized patients with COVID-19: Machine learning approach. JMIR Medical Informatics. 2021; 9(1). https://doi.org/10.2196/24207 PMID: 33400679

29. Smirnov EA, Timoshenko DM, Andrianov SN. Comparison of regularization methods for ImageNet classification with deep convolutional Neural Networks. AASRI Procedia. 2014; 6:89–94.
30. Bilgili B., Haliloğlu M., Cinel İ. Sepsis and Acute Kidney Injury. Turkish journal of anaesthesiology and reanimation. 2014; 42(6), 294–301.
31. Liu X, DuMontier C, Hu P, Liu C, Yeung W, Mao Z, et al. Clinically interpretable machine learning models for early prediction of mortality in older patients with multiple organ dysfunction syndrome: An international multicenter retrospective study. The Journals of Gerontology: Series A. 2022;
32. Kim HB, Nguyen HT, Jin Q, Tamby S, Gelaf Romer T, Sung E, et al. Computational signatures for post-cardiac arrest trajectory prediction: Importance of early physiological time series. Anaesthesia Critical Care & Pain Medicine. 2022; 41(1):101015. https://doi.org/10.1016/j.accpm.2021.101015 PMID: 34968747
33. Kwak SK, Kim JH. Statistical Data Preparation: Management of missing values and outliers. Korean Journal of Anesthesiology. 2017; 70(4):407. https://doi.org/10.4097/kjae.2017.70.4.407 PMID: 28794835
34. Dreiseitl S, Ohno-Machado L. Logistic regression and Artificial Neural Network Classification models: A methodology review. Journal of Biomedical Informatics. 2002; 35(5–6):352–9. https://doi.org/10.1016/s1532-0464(03)00034-0 PMID: 12968784
35. Taud H, Mas JF. Multilayer Perceptron (MLP). Geomatic Approaches for Modeling Land Change Scenarios. 2017;451–5.
36. Ketkar N, Moolayil J. Introduction to pytorch. Deep Learning with Python. 2021;27–91.
37. Crossentropyloss [Internet]. CrossEntropyLoss—PyTorch 1.13 documentation. [cited 2022Dec8]. Available from: https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html.
38. Pedrycz W, Chen S-M. Deep learning: Concepts and architectures. Cham: Springer; 2020.
39. Tolles J, Meurer WJ. Logistic regression. JAMA. 2016; 316(5):533.