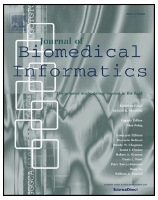
Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

FedScore: A privacy-preserving framework for federated scoring system development

Siqi Li[a], Yilin Ning[a], Marcus Eng Hock Ong[b,c,d], Bibhas Chakraborty[a,b,e,f], Chuan Hong[f], Feng Xie[a,b], Han Yuan[a], Mingxuan Liu[a], Daniel M. Buckland[g], Yong Chen[h], Nan Liu[a,b,i,*]

[a] *Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore*
[b] *Programme in Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore*
[c] *Health Services Research Centre, Singapore Health Services, Singapore, Singapore*
[d] *Department of Emergency Medicine, Singapore General Hospital, Singapore, Singapore*
[e] *Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore*
[f] *Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA*
[g] *Department of Emergency Medicine, Duke University School of Medicine, Durham, NC, USA*
[h] *Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA*
[i] *Institute of Data Science, National University of Singapore, Singapore, Singapore*

ARTICLE INFO

Keywords:
Clinical decision making
Distributed algorithm
Electronic health record
Electronic medical record
Federated learning
Scoring system

ABSTRACT

Objective: We propose FedScore, a privacy-preserving federated learning framework for scoring system generation across multiple sites to facilitate cross-institutional collaborations.

Materials and methods: The FedScore framework includes five modules: federated variable ranking, federated variable transformation, federated score derivation, federated model selection and federated model evaluation. To illustrate usage and assess FedScore's performance, we built a hypothetical global scoring system for mortality prediction within 30 days after a visit to an emergency department using 10 simulated sites divided from a tertiary hospital in Singapore. We employed a pre-existing score generator to construct 10 local scoring systems independently at each site and we also developed a scoring system using centralized data for comparison.

Results: We compared the acquired FedScore model's performance with that of other scoring models using the receiver operating characteristic (ROC) analysis. The FedScore model achieved an average area under the curve (AUC) value of 0.763 across all sites, with a standard deviation (SD) of 0.020. We also calculated the average AUC values and SDs for each local model, and the FedScore model showed promising accuracy and stability with a high average AUC value which was closest to the one of the pooled model and SD which was lower than that of most local models.

Conclusion: This study demonstrates that FedScore is a privacy-preserving scoring system generator with potentially good generalizability.

1. Introduction

Cross-institutional collaboration has gained popularity in recent years as a way to accelerate medical research and facilitate quality improvement [1]. Widespread digitization efforts in the healthcare industry enable the use of data-driven evidence for clinical prediction models [2], which can be ideally built using centralized data pooled from as many sources as possible. Some examples of cross-regional collaborations include: the Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury [3], the Genotype to Phenotype Databases [4], the Big Data in Cardiovascular Disease [5], the Ontario Prehospital Advanced Life Support [6], the Kaiser Permanente Research Bank [7] and the Pan-Asian Resuscitation Outcomes Study [8]. However, such partnerships require data sharing, which is typically laborious and time-consuming, and sometimes even impossible due to various privacy regulations [9,10], for example, the European Union General Data Protection Regulation [11].

Federated learning (FL), sometimes referred to as distributed learning or distributed algorithms, can avoid data sharing by collectively training algorithms without exchanging patient-level data [12],

[*] Corresponding author at: Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore.
E-mail address: liu.nan@duke-nus.edu.sg (N. Liu).

https://doi.org/10.1016/j.jbi.2023.104485
Received 29 March 2023; Received in revised form 8 August 2023; Accepted 31 August 2023
Available online 1 September 2023

safeguarding patients' privacy by distributing the model training to the data-owners and aggregating their results [13]. In addition to dismantling data silos, FL could also speed up the development of much-needed AI models [14]. For instance, during the COVID-19 pandemic, Dayan et al. [14] constructed a clinical outcomes prediction model across 20 institutes using FL. Luo et al. [15] studied the demographic and clinical factors that are associated with length of stay in COVID-19 patients using a lossless, one-shot FL algorithm [15]. Vaid et al. [16] also applied FL to predict mortality in hospitalized patients with COVID-19 within 7 days. There exist many applications of FL for medical image data, most of which use black box models from computer vision. Interpretable models, on the contrary, have fewer instances of FL applications despite their popularity in clinical research.

As a type of interpretable risk scoring model [17], scoring systems have been employed in practically every diagnostic area of medicine [18] since they offer quick and simple risk assessments of numerous serious medical conditions without the use of a computer [17]. Some traditional scoring systems, such as the Glasgow Coma Scale [19] first described in 1974, rely heavily on clinician's domain expertise. More data-driven methods for building scoring systems have emerged in recent years, including the Supersparse Linear Integer Model [20], which can better deal with sparsity; approximal methods that are more computationally efficient [21,22]; and interfaces that enable flexible engagement of domain expertise, like the Interval Coded Scoring [23] and the AutoScore [24].

Regardless of development strategies, scoring systems have usually been created using single-source data, limiting application at other sites if the development data has insufficient sample size or is not representative. Although it is possible to develop scoring systems on pooled data [25], the process of doing such pooling, as noted previously, is time consuming and difficult to achieve due to privacy restrictions. As a result, frameworks for building scoring systems in a federated manner are needed to overcome such difficulties. To fill this gap, we propose FedScore, a first-of-its-kind framework for building federated scoring systems across multiple sites and demonstrated its efficacy and potential generalizability with a proof-of-concept experiment using real-world data.

2. Methods

Scoring systems are linear classification models that require users to add, subtract and multiply a few numbers in order to make a prediction [17] and have been widely utilized in the field of clinical decision-making [26–28] for risk stratification due to their interpretability and transparency. They can also assist in correcting physicians' misestimations of the probability of medical outcomes, which may be rather common [29]. Users frequently take into account a model's degree of parsimony when implementing clinical models [30], which means that a model is parsimonious if it is both sparse (i.e., it uses the least amount of variables possible) and has good prediction accuracy [30]. As an example, the AutoScore framework [24] is a computational tool to conveniently create such scores using machine learning methods, and has been well received by clinicians [31,32], because it integrates domain knowledge with data driven evidence. However, regardless of the particulars of their generation of scoring systems and accounting for model interpretability, AutoScore and other similar methods only permit the development of scoring systems using one set of pooled data. To fully exploit the growing data sources and to create less biased models, we propose our FedScore framework to achieve good parsimony and interpretability for federated data, while complying with potential privacy restrictions.

2.1. FedScore framework

The FedScore framework consists of five modules: 1) federated variable ranking; 2) federated variable transformation; 3) federated score derivation; 4) federated model selection and 5) federated model evaluation. The workflow of FedScore is illustrated in Fig. 1.

(1) federated variable ranking

Variable selection is an essential step in the development of scoring systems for parsimony. In FedScore, to construct a global model across multiple sites, it is necessary to pre-identify a set of unified candidate variables. Before ranking the variables, it is recommended to check for multicollinearity among the candidate variables and remove variables when needed in order to obtain more reliable feature importance. We employed random forests for variable importance measurement, which is a well-established approach [33–37]. The variable ranking is first performed independently via random forests at each local site, and then a global variable ranking is created by rearranging variables by their weighted ranks across all K sites. Specifically, for a single variable X_m where $1 \leq m \leq P$ and P is the total number of predictors, $q_j \in N$ denotes its rank at site j, and its global ranking is obtained by mapping all values of $\sum_{j=1}^{K} w_j q_j$ for each site to the integer set $[1, P] \subset Z$. Here, w_j is the normalized weight for site j that satisfies $\sum_{j=1}^{K} w_j = 1$. The definition and following details remain the same in the manuscript for all weights introduced. The default setting for the weight is $1/K$, indicating equal weights for all sites. In addition to the default setting, a sample size-based weight can also be applied, where $w_j = S_j/S_0$ and S_j is the sample size of site j, while S_0 is the total sample size. Users may also define their own weights to accommodate specific research considerations.

(2) Federated variable transformation

The creation of categorical variables allows for the modeling of nonlinear effects [17,24], which has been widely applied [38–44] in the development of clinical scoring systems. Following this common practice, FedScore turns continuous variables into categorical variables after unified variable ranking is established. The maximum number of categories for such transformation is pre-determined (for example, choose 5 as a usual practice), and if the maximum is surpassed, categories are combined so that the requirement is met. In our study, the quantiles of continuous variables are set at $0\%, k_1\%, k_2\%, k_3\%, k_4\%$, and 100%, where the default value of k_1, k_2, k_3, k_4 are 5, 20, 80 and 95, respectively. The unified cutoff for each continuous variable is calculated by weighting the k values acquired at each site using the same weight definition as previously described for global ranking, such that the weight for each site satisfies $\sum_{j=1}^{K} w_j = 1$.

(3) Federated score derivation

Binary outcomes are common in clinical decision making and logistic regression is a prominent method used for modelling such outcomes. Federated regression models can be realized through a variety of existing FL frameworks, including both traditional engineering-based and model-agnostic frameworks like FedAvg [45] that requires multiple iterations, and statistics-based model-specific one-shot techniques [46–51] that necessitate only one round of communication. For demonstration purposes, we have employed a one-shot privacy preserving distributed algorithm called ODAL2 [46] to perform federated logistic regression, which is communication-efficient and has been demonstrated to have low bias and high statistical efficiency [46]. This algorithm utilizes information from the lead local site (data are accessible) with the first-order and second-order gradients of the likelihood function from remotes sites (data are not accessible) to construct an approximation of the global likelihood function. The global logistic regression coefficients can then be obtained by optimizing the approximate global likelihood function. Let $x_1, x_2, \cdots x_{p-1}$ denote the $p-1$ predictors, y denote a binary outcome, and the logistic regression model can be expressed as

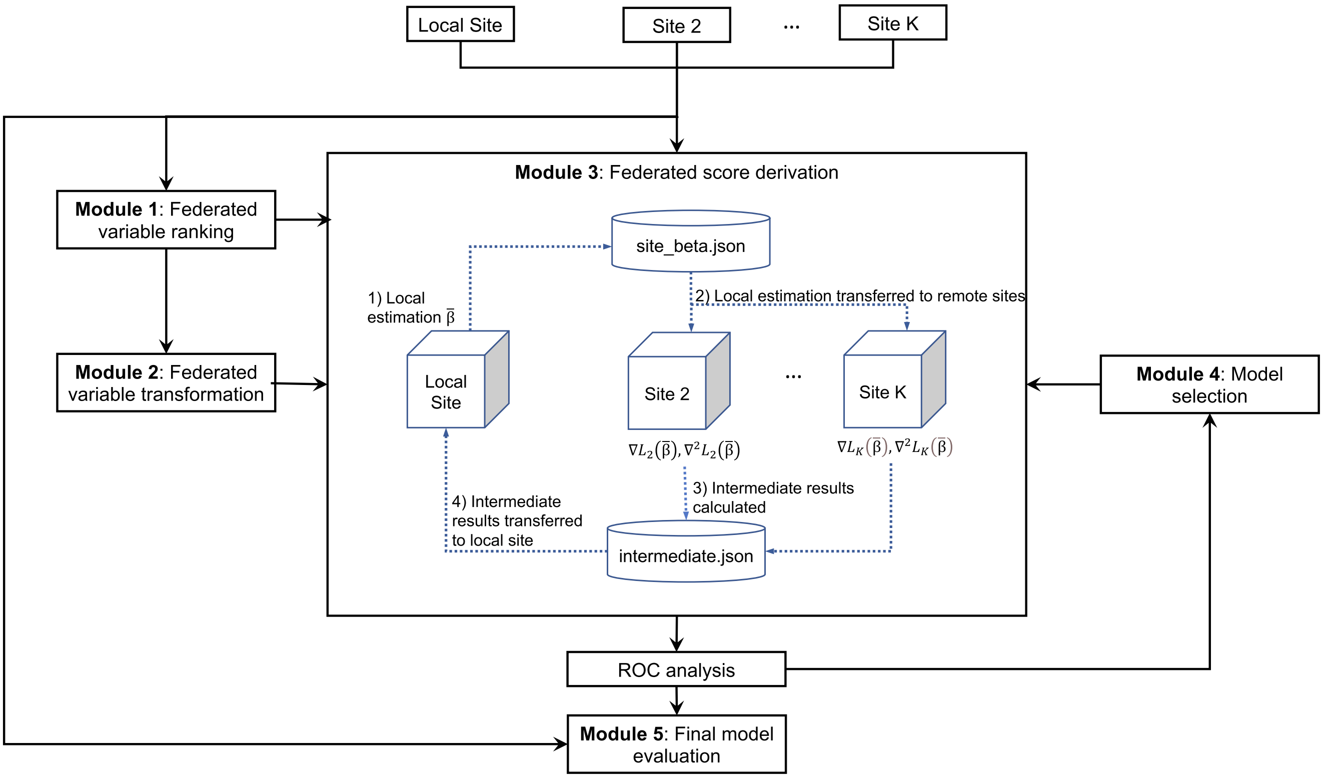

Local Site
Site 2
...
Site K
Module 1: Federated variable ranking
Module 2: Federated variable transformation
Module 3: Federated score derivation
site_beta.json
1) Local estimation $\bar{\beta}$
2) Local estimation transferred to remote sites
Local Site
Site 2
...
Site K
$\nabla L_2(\bar{\beta}), \nabla^2 L_2(\bar{\beta})$
$\nabla L_K(\bar{\beta}), \nabla^2 L_K(\bar{\beta})$
3) Intermediate results calculated
4) Intermediate results transferred to local site
intermediate.json
Module 4: Model selection
ROC analysis
Module 5: Final model evaluation

Fig. 1. Flowchart of the FedScore framework.

$$logit(Pr(y = 1|x)) = x^T\beta$$

where $x = (1, x_1, x_2, \cdots x_{p-1})^T$, β is the vector of intercept and slope coefficients, and $logit(t) = log\{t/(1-t)\}$. Suppose a total of $N = \sum_{j=1}^{K} n_j$ identically and independently distributed (i.i.d.) observations are distributed at K sites, then the log-likelihood function (LLR) of the global logistic regression by pooling data from all sites is

$$L(\beta) = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{n_j}\left[Y_{ij}x_{ij}^T\beta - log\left\{1 + exp\left(x_{ij}^T\beta\right)\right\}\right]$$

The pooled estimator $\widehat{\beta}$ can be obtained by optimizing $L(\beta)$. When data cannot be shared and the pooled likelihood function is not possible, approximation of the likelihood function is still achievable. The ODAL2 algorithm applies the idea of Taylor expansion, proposing to use first and second order gradient of LLR to perform the approximation [46]:

$$\overset{\sim 2}{L}(\beta) = L_1(\beta) + \{\nabla \mathrm{L}(\bar{\beta}) - \nabla \mathrm{L}_1(\bar{\beta})\}^T\beta +$$
$$\frac{1}{2}(\beta - \bar{\beta})^T\{\nabla^2\mathrm{L}(\bar{\beta}) - \nabla^2 \mathrm{L}_1(\bar{\beta})\}(\beta - \bar{\beta})$$

Here $\bar{\beta}$ is an initial value obtained from the regression model performed at the local site and stored for broadcasting to remote sites. $L_j(\beta) = \frac{1}{n_j}\sum_{i=1}^{n_j}\left[Y_{ij}x_{ij}^T\beta - log\left\{1 + exp\left(x_{ij}^T\beta\right)\right\}\right]$ is the LLR of the j-th site ($j = 1$ is assumed to be the local site). $\nabla L(\bar{\beta}) = \sum_{j=1}^{K} n_j \nabla L_j(\bar{\beta})/N$ is the first gradient of log-likelihood function$L(\beta)$ evaluated at initial value $\bar{\beta}$.$\nabla L_j(\bar{\beta}) = \frac{1}{n_j}\sum_{i=1}^{n_j}\left\{Y_{ij} - p_{ij}(\bar{\beta})\right\}x_{ij}$, a p-dimensional vector, is the first gradient of LLR of site j, where $p_{ij}(\bar{\beta}) = \left\{1 + exp\left(-x_{ij}^T\bar{\beta}\right)\right\}^{-1}$ and $\nabla^2 L_j(\bar{\beta}) = \frac{1}{n_j}\sum_{i=1}^{n_j} p_{ij}(\bar{\beta})\left\{1 - p_{ij}(\bar{\beta})\right\}x_{ij}x_{ij}^T$, a $p \times p$ matrix, is the second gradient of LLR of site j. Both gradients are computed at each remote site and transferred back to the local site.

Finally, the global beta estimator of β is obtained by optimizing the surrogate likelihood function. This process for constructing the global model is one-shot [46] as illustrated in Fig. 1, and neither of the shared files contain any patient level information, which guarantees privacy. Federated scores are obtained by having coefficients in the global logistic regression model rounded to integers and mapped to interval $[0, S_{max}]$, where S_{max} is the maximum score pre-decided by users, e.g., 100.

(4) Federated model selection

Model selection is performed using parsimony plots generated on validation data, with variables added incrementally based on the variable ranking for the x-axis and AUC values for the y-axis. A general model selection criteria could be defined by maximizing $\Psi_\mathrm{m} = \sum w_j\phi_j(p_1, p_2, p_3, \cdots p_m)$, where w_j is the weight for site j as previously described for global ranking, ϕ_j measures a score's performance on the jth validation set (e.g. AUC value) and m is a pre-specified number of total variables to include, which should be uniform across all sites. Different constraints can be added for the optimization task. For example, the total number of variables m may not exceed an integer number D. The set of variables $\{p_1, p_2, \cdots p_m\}$ may also be set to satisfy certain subjective standard required by users. For instance, users may decide (based on domain knowledge) that a set of variables $\{x_1, x_2, .. x_q\}$, where $q \leq m$ must be included in the final scoring system regardless of the results provided by variable important analysis. Moreover, Ψ may be maximized using a number of d of variables that is smaller than m, as long as increasing the number of variables from d to m has little impact on the change in Ψ: $|\Psi_m - \Psi_d| \leq \epsilon$, where the size of ∊ may be decided intuitively by users based on parsimony plots.

After final variables are confirmed based on the selected model, a new model is refitted via module 2) so that the final model is as parsimony as possible.

(5) Federated model evaluation

The performance of the final model is validated on each site engaged in the FedScore framework. Following the Ψ_m defined in step 4), the overall weighted performance of a federated score is $M_1 = \sum w_j \mu_j(p_1, p_2, p_3, \cdots p_m)$, where μ_j is the score's performance on j th testing set and w_j is the same weight as previously defined; and $M_2 = \left(\sum w_j\left(M_1 - \mu_j\right)^2\right)^{1/2}$ is a measurement of weighted performance variation across sites. A higher M_1 value and lower M_2 value indicate a score's better performance and generalizability.

The FedScore framework has been implemented in R 4.0.3 and code is available at https://github.com/nliulab/FedScore.

2.2. Experiment

The initial study cohort was formed by selecting emergency department (ED) visits in 2016 and 2017, using the EHR data of Singapore General Hospital (SGH) extracted from the SingHealth Electronic Health Intelligence System. A waiver of consent was granted for EHR data collection and retrospective analysis, and the study has been approved by the Singapore Health Services' Centralized Institutional Review Board, with all data deidentified. After excluding patients under the age of 18 and those with missing values, the remaining cohort was randomly divided into 10 sites for demonstration purpose, in the proportion of 4%, 5%, 7%, 9%, 10%, 11%, 12%, 13%, 14%, and 15% respectively. Fig. 2 depicts the process of cohort formation.

The outcome in this study was whether a patient died within 30 days after ED admissions. Candidate variables were determined based on a recent work [31], the study cohort for which was also obtained from SGH ED data. The candidate predictors include a total of 29 variables in 5 categories: (1) demographics information: age, sex and race; (2) PACS [52] triage categories (P1, P2, P3 and P4), shift time (8 AM to 4 PM, 4 PM to midnight, Midnight to 8 AM), and day of week (Friday, Monday, Weekend, Midweek); (3) vital signs: pulse (beats/min), respiration (times/min), peripheral capillary oxygen saturation (SpO_2; %), diastolic blood pressure (mm Hg), and systolic blood pressure (mm Hg); (4) comorbidities: myocardial infarction, congestive heart failure, peripheral vascular disease, stroke, dementia, chronic pulmonary disease, rheumatoid disease, peptic ulcer disease, diabetes, hemiplegia or paraplegia, kidney disease, and liver disease; (5) previous health care usage: ED visits in the past year, surgical procedures in the past year, ICU admissions in the past year, and high-dependency admissions in the past year.

We first used the variance inflation factor to detect variable correlation and found weak multicollinearity in the data. We then employed AutoScore to create baseline models for local and pooled comparisons with our FedScore framework. Three groups of analysis were performed: 1) 10 local scores trained independently on each site with AutoScore; 2) one federated score trained using all sites without data sharing with FedScore; 3) one pooled score generated using centralized data with AutoScore, which is the ideal case but usually impossible in most real world settings. All models were chosen based on corresponding parsimony plots, with a predefined criterion that the maximum number of variables in a model should not exceed 8 and adding more variables until there was no significant improvement in AUC. In order to perform straightforward comparisons, the cutoffs and weights used during scoring system development were default options specified in Section 2.1 and all processes involved were data-driven without refining, which engaged expert knowledge from clinical practice.

3. Results

A total of 80,613 individual ED admission episodes were randomly divided into 10 sites, with sample size ranging from 3224 to 12,092 and the training, validation and testing sets of each site were obtained by

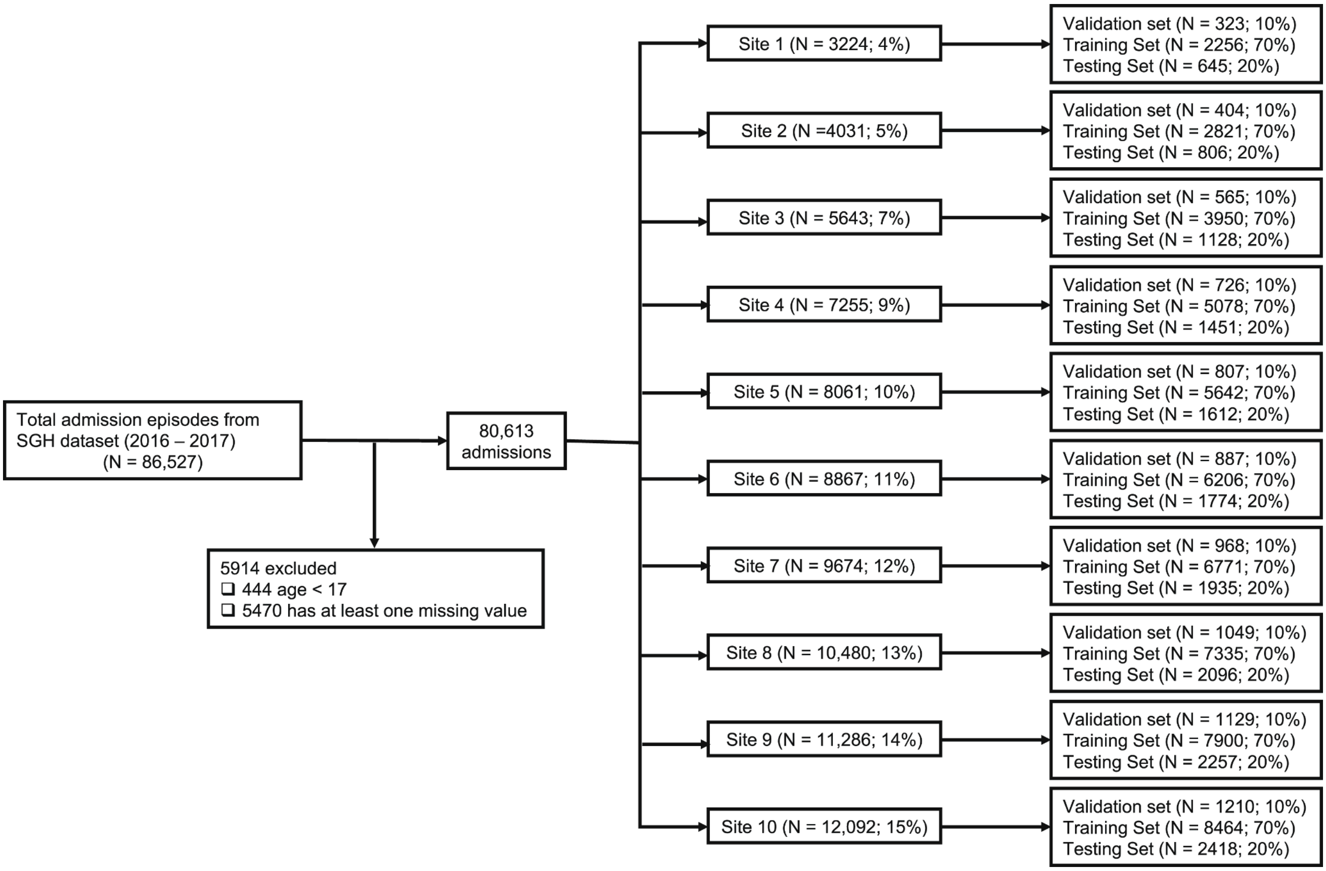

Total admission episodes from SGH dataset (2016 – 2017) (N = 86,527)
80,613 admissions
5914 excluded
❑ 444 age < 17
❑ 5470 has at least one missing value
Site 1 (N = 3224; 4%)
Site 2 (N =4031; 5%)
Site 3 (N = 5643; 7%)
Site 4 (N = 7255; 9%)
Site 5 (N = 8061; 10%)
Site 6 (N = 8867; 11%)
Site 7 (N = 9674; 12%)
Site 8 (N = 10,480; 13%)
Site 9 (N = 11,286; 14%)
Site 10 (N = 12,092; 15%)
Validation set (N = 323; 10%) Training Set (N = 2256; 70%) Testing Set (N = 645; 20%)
Validation set (N = 404; 10%) Training Set (N = 2821; 70%) Testing Set (N = 806; 20%)
Validation set (N = 565; 10%) Training Set (N = 3950; 70%) Testing Set (N = 1128; 20%)
Validation set (N = 726; 10%) Training Set (N = 5078; 70%) Testing Set (N = 1451; 20%)
Validation set (N = 807; 10%) Training Set (N = 5642; 70%) Testing Set (N = 1612; 20%)
Validation set (N = 887; 10%) Training Set (N = 6206; 70%) Testing Set (N = 1774; 20%)
Validation set (N = 968; 10%) Training Set (N = 6771; 70%) Testing Set (N = 1935; 20%)
Validation set (N = 1049; 10%) Training Set (N = 7335; 70%) Testing Set (N = 2096; 20%)
Validation set (N = 1129; 10%) Training Set (N = 7900; 70%) Testing Set (N = 2257; 20%)
Validation set (N = 1210; 10%) Training Set (N = 8464; 70%) Testing Set (N = 2418; 20%)

Fig. 2. Flowchart of the study cohorts' formation. SGH: Singapore General Hospital.

randomly splitting at ratios of 70%, 10%, and 20% respectively, as shown in Fig. 2. A comprehensive summary of the baseline characteristics of the overall cohort and each participating site can be found in eTable 1.

We compared the performance of the federated score developed by FedScore with the pooled score developed using all data and the 10 local scores independently developed at each site by AutoScore. Fig. 3 depicts how each score performed on the testing datasets of each site, with twelve subplots.

For each subplot, a scoring model's performance on each of the ten sites is presented in horizontal lines using the corresponding AUC values and its 95% confidence intervals (CIs). The vertical edges of the grey rectangular frame in each subplot reflect the mean of ten AUC values plus/minus their standard deviation (SD) and as a result, the width of each grey rectangular frame represents the degree of performance variation of a model across all sites. The detailed AUC values, CIs and SDs are reported in Supplementary eTable 2. The ROC curves of each model on each site's testing data is provided in eFigure 1 of the Supplementary Materials. The scoring tables for each model and corresponding parsimony plots were also provided in eTable 3 and eFigure 1 of the Supplementary Materials.

With the information presented in Fig. 3 and eTable 2, we summarize the following main observations: 1) the federated score achieved good performance, with an average AUC value across all sites of 0.763, better than that of the local models and close to the one of the pooled model; 2) the AUC variance of federated score is among the smallest ones, and although the SDs for local models of site 2 and 5 appear to be slightly smaller, their averaged AUC values are lower; 3) the performance of the federated score on some sites are better than the model developed locally at that site (e.g. 0.7804 > 0.7300 at site 7).

4. Discussion

FedScore is among the first frameworks that aims to generalize unified scores across multiple sites while preserving privacy. The scalable and adaptable architecture offers potential solutions for improving model generalizability and stability across isolated clinical datasets.

Whereas scoring systems have been widely utilized in clinical domains, few existing FL applications have focused on them despite their prevalence, reflecting the phenomenon that existing biomedical FL applications have a tendency to favor black box models [53] over more interpretable ML models. To meet physicians' expectations for model simplicity and transparency, FL applications of interpretable models require more customization and modification compared to black box

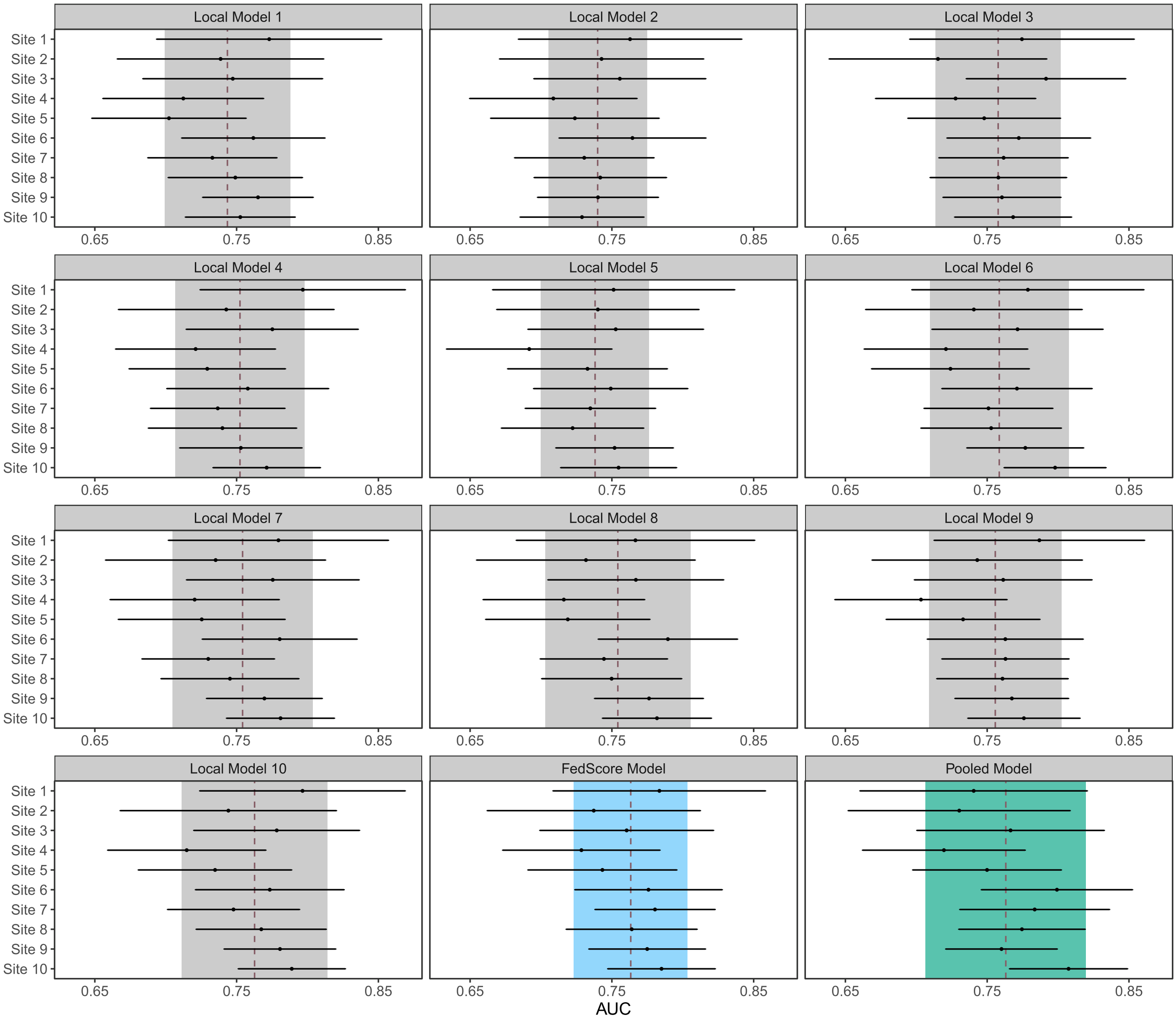


Fig. 3. Comparison of FedScore performance with baselines using AUC and 95% CI of the FedScore model and other models applied to each of the 10 sites.

model implementations with well-established FL frameworks available from the computer science community. A simple and straightforward scoring table with a lower AUC value for risk stratification, for example, would be preferred by clinicians over a black box model with a higher AUC value in the ED. As a result, more cautious designs for FL applications of interpretable models are required and FedScore deals with this issue by emphasizing model parsimony and enabling flexible process monitoring for users. Future FL applications in clinical sciences should take similar factors into account if the research questions favor transparent solutions rather than merely being concerned with model performance.

FL studies in the biomedical field differ from the ones in computer science, albeit sharing similar origins. In many standard engineering FL contexts, since a single client cannot create models independently, attention has been paid to technical details such as data partitioning schemes and various privacy mechanisms [54]. In clinical domains however, data are frequently formed at the hospital or institution level, making local models feasible in these cases. Under these circumstances, generalizability (models' ability to generalize their performance to a new setting [55]) and stability of global models relative to local models become more crucial, but these factors are not sufficiently considered in many existing FL frameworks that are being developed. The results in section 3 show that by a co-training process via FL, a global model prediction framework such as FedScore can achieve less variation than locally developed ones while still maintaining good performance. This benefit of FL is promising for medical research that seeks dependable high risk decision making.

Data constraints, such as biased data and small datasets are considered a source of ML misuse [56], yet investigating such misconduct is not as feasible as developing models. Despite the emphasis [55] on external validations, less than 10% of clinical prediction studies reported to have done so [57]. Instead of training a model on single site and subsequently testing and modifying it on other sites, constructing a model with sufficient and representative data through privacy-preserving means may be a more viable solution. FedScore and its future extensions could potentially aid in reducing model inconsistency across cohorts, leading to more trustworthy decision-making for medical research.

Although we have only used one binary outcome example for illustration, our FedScore framework is scalable and versatile, given that modules could be appropriately modified to accommodate different clinical research questions. For instance, the score derivation module could be modified to accommodate survival or ordinal outcomes, and additional privacy-preserving FL frameworks and topologies might also be added to offer more options. We anticipate that FedScore and its future extensions could together act as some foundations for creating more trustworthy clinical scoring systems in approaches that safeguard data privacy.

Limitations

Results were obtained from homogenous data split from a single source without consideration of site-specific real-world heterogeneity. FedScore may encounter problems with heterogeneous medical data because the current ODAL2 algorithm in module 3 requires that the data across different sites are homogeneous, similar to the majority of the FL and distributed methods currently in use [12,58]. Despite the fact that FedScore does not currently address data heterogeneity, its scalability allows for continuous updates with cutting-edge solutions for better handling of the issue. We anticipate that the overall functionality and applicability of FedScore in various clinical research settings will be significantly improved by this ongoing process of improvement.

Future work

Our future work will involve international collaboration to develop FedScore with more heterogeneous datasets. We plan to extend FedScore by incorporating the two state-of-art FL algorithms that account for the between-site heterogeneity. The first strategy is to use the dCLR algorithm [59], motivated from a novel pairwise conditional logistic regression, to estimate the common regression coefficients and then estimate the site-specific intercept locally for each site. The second strategy is to adopt the lossless, few-shot dPQL algorithm [60], which has been used to rank the performance of different hospitals while considering the case-mix situation across sites (i.e., different hospitals are treating different patients).

Future research could also explore the use of site-personalized federated models rather than a uniform model across all sites. One limitation of the current FedScore version is the requirement for uniform cutoffs in the federated model, which may not be feasible if clinical practices vary significantly among sites, and clinicians prefer more personalized models. To address this, alternative strategies such as personalized federated learning [61] or domain adaptation [62] could be considered. Additionally, incorporating common data models like OMOP [63], I2B2 [64], Mini-Sentinel [65] and 4CE [66] at the data-preprocessing stage could enhance the integration of health data from heterogeneous sources and enable systematic analysis [67], which is often overlooked in existing FL applications in healthcare [68]. The scalable nature of FedScore provides a platform for the incorporation of such enhancements, but further research is necessary to fully assess their feasibility and overall impact.

5. Conclusion

We have proposed FedScore, a privacy-preserving scoring systems and used a 30-day mortality prediction task to show proof-of-concept. We have demonstrated its potential to build effective federated clinical scores that are more generalizable, with lower performance variability across sites. FedScore is a first-of-its-kind framework for constructing scoring systems based on distributed algorithms, bridging a gap in current medical research. While demonstrated for binary outcomes, the application of FedScore can be extended for settings with other types of clinical outcomes and greater heterogeneity across sites with future developments in FL and clinical prediction methods, enabling its use in a wide range of different medical contexts.

6. Funding

This work was supported by the Duke/Duke-NUS Collaboration grant. The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

7. Code availability

The code for FedScore is available at: https://github.com/nliulab/FedScore.

CRediT authorship contribution statement

Siqi Li: Conceptualization, Data curation, Software, Formal analysis, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Yilin Ning:** Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Marcus Eng Hock Ong:** Validation, Investigation, Methodology, Writing – review & editing. **Bibhas Chakraborty:** Validation, Investigation, Writing – review & editing. **Chuan Hong:** Validation, Investigation, Writing – review & editing. **Feng Xie:** Validation, Investigation, Writing – review & editing. **Han Yuan:** Validation, Investigation, Writing – review & editing. **Mingxuan Liu:** Validation, Investigation, Writing – review & editing. **Daniel M. Buckland:** Validation, Investigation, Writing – review & editing. **Yong Chen:** Validation, Investigation, Writing – review & editing. **Nan Liu:**

Conceptualization, Investigation, Methodology, Project administration, Funding acquisition, Resources, Supervision.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: NL, SL and MEHO hold a patent related to the federated scoring system. The other authors declare no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104485.

References

[1] T.-T. Kuo, A. Pham, Detecting model misconducts in decentralized healthcare federated learning, Int. J. Med. Inf. 158 (2022), 104658, https://doi.org/10.1016/j.ijmedinf.2021.104658.

[2] D. Gotz, D. Borland, Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization, IEEE Comput. Graph. Appl. 36 (2016) 90–96, https://doi.org/10.1109/MCG.2016.59.

[3] A.I.R. Maas, D.K. Menon, E.W. Steyerberg, et al., Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury (CENTER-TBI): a prospective longitudinal observational study, Neurosurgery 76 (2015) 67–80, https://doi.org/10.1227/NEU.0000000000000575.

[4] A.J. Webb, G.A. Thorisson, A.J. Brookes, et al., An informatics project and online "Knowledge Centre" supporting modern genotype-to-phenotype research, Hum. Mutat. 32 (2011) 543–550, https://doi.org/10.1002/humu.21469.

[5] S. Anker, F.W. Asselbergs, G. Brobert, et al., Big Data in Cardiovascular Disease, Eur. Heart J. 38 (2017) 1863–1865, https://doi.org/10.1093/eurheartj/ehx283.

[6] I.G. Stiell, G.A. Wells, V.J. DeMaio, et al., Modifiable Factors Associated With Improved Cardiac Arrest Survival in a Multicenter Basic Life Support/Defibrillation System: OPALS Study Phase I Results, Ann. Emerg. Med. 33 (1999) 44–50, https://doi.org/10.1016/S0196-0644(99)70415-4.

[7] Kaiser Permanente Research Bank - Kaiser Permanente. Kais. Perm. Res. Bank. https://researchbank.kaiserpermanente.org/ (accessed 10 Aug 2022).

[8] M.E.H. Ong, S.D. Shin, H. Tanaka, et al., Pan-Asian Resuscitation Outcomes Study (PAROS): rationale, methodology, and implementation, Acad. Emerg. Med. 18 (2011) 890–897, https://doi.org/10.1111/j.1553-2712.2011.01132.x.

[9] R.S. Antunes, C. André da Costa, A. Küderle, et al., Federated Learning for Healthcare: Systematic Review and Architecture Proposal, ACM Trans. Intell. Syst. Technol. 13 (54) (2022), https://doi.org/10.1145/3501813.

[10] D.C. Nguyen, Q.-V. Pham, P.N. Pathirana, et al., Federated Learning for Smart Healthcare: A Survey, ACM Comput. Surv. 55 (60) (2022), https://doi.org/10.1145/3501296.

[11] C.J. Hoofnagle, B. van der Sloot, F.Z. Borgesius, The European Union general data protection regulation: what it is and what it means, Inf. Commun. Technol. Law 28 (2019) 65–98, https://doi.org/10.1080/13600834.2019.1573501.

[12] N. Rieke, J. Hancox, W. Li, et al., The future of digital health with federated learning, Npj Digit Med. 3 (2020) 1–7, https://doi.org/10.1038/s41746-020-00323-1.

[13] M.J. Sheller, B. Edwards, G.A. Reina, et al., Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, Sci. Rep. 10 (2020) 12598, https://doi.org/10.1038/s41598-020-69250-1.

[14] I. Dayan, H.R. Roth, A. Zhong, et al., Federated learning for predicting clinical outcomes in patients with COVID-19, Nat. Med. 27 (2021) 1735–1743, https://doi.org/10.1038/s41591-021-01506-3.

[15] C. Luo, M.N. Islam, N.E. Sheils, et al., DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models, Nat. Commun. 13 (2022) 1678, https://doi.org/10.1038/s41467-022-29160-4.

[16] A. Vaid, S.K. Jaladanki, J. Xu, et al., Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach, JMIR Med. Inform. 9 (2021) e24207.

[17] C. Rudin, C. Chen, Z. Chen, et al., Interpretable machine learning: Fundamental principles and 10 grand challenges, Stat. Surv. 16 (2022) 1–85, https://doi.org/10.1214/21-SS133.

[18] V. Fleig, F. Brenck, M. Wolff, et al., Scoring systems in intensive care medicine : principles, models, application and limits, Anaesthesist 60 (2011) 963–974, https://doi.org/10.1007/s00101-011-1942-8.

[19] ASSESSMENT OF COMA AND IMPAIRED CONSCIOUSNESS - The Lancet. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(74)91639-0/fulltext (accessed 6 Jun 2022).

[20] B. Ustun, C. Rudin, Supersparse linear integer models for optimized medical scoring systems, Mach. Learn. 102 (2016) 349–391, https://doi.org/10.1007/s10994-015-5528-6.

[21] N. Sokolovska, Y. Chevaleyre, K. Clément, et al., The fused lasso penalty for learning interpretable medical scoring systems, in: 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 4504–11. doi:10.1109/IJCNN.2017.7966427.

[22] N. Sokolovska, Y. Chevaleyre, J.-D. Zucker, A Provable Algorithm for Learning Interpretable Scoring Systems, in: Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics. PMLR 2018. 566–74.https://proceedings.mlr.press/v84/sokolovska18a.html (accessed 8 Aug 2022).

[23] L. Billiet, S.V. Huffel, V.V. Belle, Interval Coded Scoring: a toolbox for interpretable scoring systems, PeerJ Comput. Sci. 4 (2018) e150.

[24] F. Xie, B. Chakraborty, M.E.H. Ong, et al., AutoScore: A Machine Learning-Based Automatic Clinical Score Generator and Its Application to Mortality Prediction Using Electronic Health Records, JMIR Med. Inform. 8 (2020) e21798.

[25] N. Liu, M. Liu, X. Chen, et al., Development and validation of an interpretable prehospital return of spontaneous circulation (P-ROSC) score for patients with out-of-hospital cardiac arrest using machine learning: A retrospective study, eClinicalMedicine 48 (2022), 101422, https://doi.org/10.1016/j.eclinm.2022.101422.

[26] M.M. Churpek, T.C. Yuen, S.Y. Park, et al., Derivation of a cardiac arrest prediction model using ward vital signs*, Crit. Care Med. 40 (2012) 2102–2108, https://doi.org/10.1097/CCM.0b013e318250aa5a.

[27] G.B. Smith, D.R. Prytherch, P. Meredith, et al., The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death, Resuscitation 84 (2013) 465–470, https://doi.org/10.1016/j.resuscitation.2012.12.016.

[28] W. Brady, K. de Souza, The HEART score: A guide to its application in the emergency department, Turk. J. Emerg. Med. 18 (2018) 47–51, https://doi.org/10.1016/j.tjem.2018.04.004.

[29] H.R. Arkes, S.K. Aberegg, K.A. Arpin, Analysis of Physicians' Probability Estimates of a Medical Outcome Based on a Sequence of Events, JAMA Netw. Open 5 (2022) e2218804.

[30] L.N. Sanchez-Pinto, L.R. Venable, J. Fahrenbach, et al., Comparison of variable selection methods for clinical predictive modeling, Int. J. Med. Inf. 116 (2018) 10–17, https://doi.org/10.1016/j.ijmedinf.2018.05.006.

[31] F. Xie, M.E.H. Ong, J.N.M.H. Liew, et al., Development and Assessment of an Interpretable Machine Learning Triage Tool for Estimating Mortality After Emergency Admissions, JAMA Netw. Open 4 (2021) e2118467.

[32] Y. Ang, S. Li, M.E.H. Ong, et al., Development and validation of an interpretable clinical score for early identification of acute kidney injury at the emergency department, Sci. Rep. 12 (2022) 7111, https://doi.org/10.1038/s41598-022-11129-4.

[33] B. Gregorutti, B. Michel, P. Saint-Pierre, Correlation and variable importance in random forests, Stat. Comput. 27 (2017) 659–678, https://doi.org/10.1007/s11222-016-9646-1.

[34] E.V.A. Sylvester, P. Bentzen, I.R. Bradbury, et al., Applications of random forest feature selection for fine-scale genetic population assignment, Evol. Appl. 11 (2018) 153–165, https://doi.org/10.1111/eva.12524.

[35] J.K. Jaiswal, R. Samikannu, Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression, in: 2017 World Congress on Computing and Communication Technologies (WCCCT), 2017, pp. 65–8. doi:10.1109/WCCCT.2016.25.

[36] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Variable selection using random forests, Pattern Recogn. Lett. 31 (2010) 2225–2236, https://doi.org/10.1016/j.patrec.2010.03.014.

[37] A.P. Marques Ramos, L. Prado Osco, D. Elis Garcia Furuya, et al., A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices, Comput. Electron Agric. 178 (2020) 105791, https://doi.org/10.1016/j.compag.2020.105791.

[38] L.G. Forni, T. Dawes, H. Sinclair, et al., Identifying the patient at risk of acute kidney injury: a predictive scoring system for the development of acute kidney injury in acute medical patients, Nephron Clin. Pract. 123 (2013) 143–150, https://doi.org/10.1159/000351509.

[39] M.E. Charlson, P. Pompei, K.L. Ales, et al., A new method of classifying prognostic comorbidity in longitudinal studies: development and validation, J. Chronic Dis. 40 (1987) 373–383, https://doi.org/10.1016/0021-9681(87)90171-8.

[40] J.L. Vincent, R. Moreno, J. Takala, et al., The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine, Intensive Care Med. 22 (1996) 707–710, https://doi.org/10.1007/BF01709751.

[41] A.J. Six, B.E. Backus, J.C. Kelder, Chest pain in the emergency room: value of the HEART score, Neth Heart J 16 (2008) 191–196.

[42] M. Jones, NEWSDIG: The National Early Warning Score Development and Implementation Group, Clin. Med. 12 (2012) 501–503, https://doi.org/10.7861/clinmedicine.12-6-501.

[43] E. Seth, E.M. BentleyEllison, et al., The SPOTS System: An Ocular Scoring System Optimized for Use in Modern Preclinical Drug Development and Toxicology, J. Ocul. Pharmacol. Ther. Published Online First 1 (December 2017), https://doi.org/10.1089/jop.2017.0108.

[44] E. Baldi, M.L. Caputo, S. Savastano, et al., An Utstein-based model score to predict survival to hospital admission: The UB-ROSC score, Int. J. Cardiol. 308 (2020) 84–89, https://doi.org/10.1016/j.ijcard.2020.01.032.

[45] McMahan Brendan, E. Moore, D. Ramage, et al., Communication-Efficient Learning of Deep Networks from Decentralized Data, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR 2017, pp. 1273–82. https://proceedings.mlr.press/v54/mcmahan17a.html (accessed 5 Jul 2022).

[46] R. Duan, M.R. Boland, Z. Liu, et al., Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed

algorithm, J. Am. Med. Inform. Assoc. 27 (2019) 376–385, https://doi.org/10.1093/jamia/ocz199.

[47] R. Duan, M.R. Boland, J.H. Moore, et al., ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites, Pac. Symp. Biocomput. 24 (2019) 30–41.

[48] R. Duan, C. Luo, M.J. Schuemie, et al., Learning from local to global: An efficient distributed algorithm for modeling time-to-event data, J. Am. Med. Inform. Assoc. 27 (2020) 1028–1036, https://doi.org/10.1093/jamia/ocaa044.

[49] M.J. Edmondson, C. Luo, Md Nazmul Islam, et al., Distributed Quasi-Poisson regression algorithm for modeling multi-site count outcomes in distributed data networks, J. Biomed. Inform. 131 (2022), https://doi.org/10.1016/j.jbi.2022.104097.

[50] X. Wang, H.G. Zhang, X. Xiong, et al., SurvMaximin: Robust federated approach to transporting survival risk prediction models, J. Biomed. Inform. 134 (2022), 104176, https://doi.org/10.1016/j.jbi.2022.104176.

[51] M.J. Edmondson, C. Luo, R. Duan, et al., An efficient and accurate distributed learning algorithm for modeling multi-site zero-inflated count outcomes, Sci. Rep. 11 (2021) 19647, https://doi.org/10.1038/s41598-021-99078-2.

[52] R.Y. Fong, W.S.S. Glen, A.K. Mohamed Jamil, et al., Comparison of the Emergency Severity Index versus the Patient Acuity Category Scale in an emergency setting, Int. Emerg. Nurs. 41 (2018) 13–18, https://doi.org/10.1016/j.ienj.2018.05.001.

[53] M.G. Crowson, D. Moukheiber, A.R. Arévalo, et al., A systematic review of federated learning applications for biomedical data, PLOS Digit Health 1 (2022) e0000033.

[54] C. Zhang, Y. Xie, H. Bai, et al., A survey on federated learning, Knowl.-Based Syst. 216 (2021), 106775, https://doi.org/10.1016/j.knosys.2021.106775.

[55] A.A.H. de Hond, A.M. Leeuwenberg, L. Hooft, et al., Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review, Npj Digit Med 5 (2022) 1–13, https://doi.org/10.1038/s41746-021-00549-7.

[56] V. Volovici, N.L. Syn, A. Ercole, et al., Steps to avoid overuse and misuse of machine learning in clinical research, Nat. Med. (2022) 1–4, https://doi.org/10.1038/s41591-022-01961-6.

[57] C. Yang, J.A. Kors, S. Ioannou, et al., Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review, J. Am. Med. Inform. Assoc. 29 (2022) 983–989, https://doi.org/10.1093/jamia/ocac002.

[58] Federated Learning: Challenges, Methods, and Future Directions. http://ieeexplore.ieee.org/document/9084352 (accessed 23 Jun 2022).

[59] J. Tong, C. Luo, M.N. Islam, et al., Distributed learning for heterogeneous clinical data with application to integrating COVID-19 data across 230 sites, Npj Digit Med. 5 (2022) 1–8, https://doi.org/10.1038/s41746-022-00615-8.

[60] C. Luo, M.N. Islam, N.E. Sheils, et al., dPQL: a lossless distributed algorithm for generalized linear mixed model with application to privacy-preserving hospital profiling, J. Am. Med. Inform. Assoc. 29 (2022) 1366–1371, https://doi.org/10.1093/jamia/ocac067.

[61] A. Fallah, A. Mokhtari, A. Ozdaglar, Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach, in: Advances in Neural Information Processing Systems. Curran Associates, Inc. 2020, pp. 3557–68. https://proceedings.neurips.cc/paper/2020/hash/24389bfe4fe2eba8bf9aa9203a44cdad-Abstract.html (accessed 4 Jan 2023).

[62] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J Big Data 3 (2016) 9, https://doi.org/10.1186/s40537-016-0043-6.

[63] E.A. Voss, R. Makadia, A. Matcho, et al., Feasibility and utility of applications of the common data model to multiple, disparate observational health databases, J. Am. Med. Inform. Assoc. JAMIA 22 (2015) 553–564, https://doi.org/10.1093/jamia/ocu023.

[64] S.N. Murphy, G. Weber, M. Mendis, et al., Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), J. Am. Med. Inform. Assoc. 17 (2010) 124–130, https://doi.org/10.1136/jamia.2009.000893.

[65] R.E. Behrman, J.S. Benner, J.S. Brown, et al., Developing the Sentinel System — A National Resource for Evidence Development, N. Engl. J. Med. 364 (2011) 498–499, https://doi.org/10.1056/NEJMp1014427.

[66] G.A. Brat, G.M. Weber, N. Gehlenborg, et al., International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium, NPJ Digit Med. 3 (2020) 109, https://doi.org/10.1038/s41746-020-00308-0.

[67] S. Kohler, D. Boscá, F. Kärcher, et al., Eos and OMOCL: Towards a seamless integration of openEHR records into the OMOP Common Data Model, J. Biomed. Inform. (2023), https://doi.org/10.1016/j.jbi.2023.104437.

[68] S. Li, P. Liu, G.G. Nascimento, et al., Federated and distributed learning applications for electronic health records and structured medical data: A scoping review, J. Am. Med. Inform. Assoc. (2023). https://doi.org/10.1093/jamia/ocad170.