A scalable federated learning solution for secondary care using low-cost microcomputing: privacy-preserving development and evaluation of a COVID-19 screening test in UK hospitals

Andrew A S Soltan, Anshul Thakur, Jenny Yang, Anoop Chauhan, Leon G D'Cruz, Phillip Dickson, Marina A Soltan, David R Thickett, David W Eyre, Tingting Zhu, David A Clifton

Summary

Background Multicentre training could reduce biases in medical artificial intelligence (AI); however, ethical, legal, and technical considerations can constrain the ability of hospitals to share data. Federated learning enables institutions to participate in algorithm development while retaining custody of their data but uptake in hospitals has been limited, possibly as deployment requires specialist software and technical expertise at each site. We previously developed an artificial intelligence-driven screening test for COVID-19 in emergency departments, known as CURIAL-Lab, which uses vital signs and blood tests that are routinely available within 1 h of a patient's arrival. Here we aimed to federate our COVID-19 screening test by developing an easy-to-use embedded system—which we introduce as full-stack federated learning—to train and evaluate machine learning models across four UK hospital groups without centralising patient data.

Methods We supplied a Raspberry Pi 4 Model B preloaded with our federated learning software pipeline to four National Health Service (NHS) hospital groups in the UK: Oxford University Hospitals NHS Foundation Trust (OUH; through the locally linked research University, University of Oxford), University Hospitals Birmingham NHS Foundation Trust (UHB), Bedfordshire Hospitals NHS Foundation Trust (BH), and Portsmouth Hospitals University NHS Trust (PUH). OUH, PUH, and UHB participated in federated training, training a deep neural network and logistic regressor over 150 rounds to form and calibrate a global model to predict COVID-19 status, using clinical data from patients admitted before the pandemic (COVID-19-negative) and testing positive for COVID-19 during the first wave of the pandemic. We conducted a federated evaluation of the global model for admissions during the second wave of the pandemic at OUH, PUH, and externally at BH. For OUH and PUH, we additionally performed local fine-tuning of the global model using the sites' individual training data, forming a site-tuned model, and evaluated the resultant model for admissions during the second wave of the pandemic. This study included data collected between Dec 1, 2018, and March 1, 2021; the exact date ranges used varied by site. The primary outcome was overall model performance, measured as the area under the receiver operating characteristic curve (AUROC). Removable micro secure digital (microSD) storage was destroyed on study completion.

Findings Clinical data from 130 941 patients (1772 COVID-19-positive), routinely collected across three hospital groups (OUH, PUH, and UHB), were included in federated training. The evaluation step included data from 32 986 patients (3549 COVID-19-positive) attending OUH, PUH, or BH during the second wave of the pandemic. Federated training of a global deep neural network classifier improved upon performance of models trained locally in terms of AUROC by a mean of 27·6% (SD 2·2): AUROC increased from 0·574 (95% CI 0·560–0·589) at OUH and 0·622 (0·608–0·637) at PUH using the locally trained models to 0·872 (0·862–0·882) at OUH and 0·876 (0·865–0·886) at PUH using the federated global model. Performance improvement was smaller for a logistic regression model, with a mean increase in AUROC of 13·9% (0·5%). During federated external evaluation at BH, AUROC for the global deep neural network model was 0·917 (0·893–0·942), with 89·7% sensitivity (83·6–93·6) and 76·6% specificity (73·9–79·1). Site-specific tuning of the global model did not significantly improve performance (change in AUROC <0·01).

Interpretation We developed an embedded system for federated learning, using microcomputing to optimise for ease of deployment. We deployed full-stack federated learning across four UK hospital groups to develop a COVID-19 screening test without centralising patient data. Federation improved model performance, and the resultant global models were generalisable. Full-stack federated learning could enable hospitals to contribute to AI development at low cost and without specialist technical expertise at each site.

Funding The Wellcome Trust, University of Oxford Medical and Life Sciences Translational Fund.

Lancet Digit Health 2024; 6: e93–104

See **Comment** page e88

Oxford University Hospitals NHS Foundation Trust, Oxford, UK (A A S Soltan MRCP, Prof D W Eyre DPhil); **Department of Oncology** (A A S Soltan), **Institute of Biomedical Engineering, Department of Engineering Science** (A A S Soltan, A Thakur PhD, J Yang MSc, Prof T Zhu DPhil, Prof D A Clifton DPhil), **Big Data Institute, Nuffield Department of Population Health** (A A S Soltan, Prof D W Eyre), **and Nuffield Department of Primary Care Health Sciences** (A A S Soltan), **University of Oxford, Oxford, UK; Portsmouth Hospitals University NHS Trust, Portsmouth, UK** (Prof A Chauhan FRCP, L G D'Cruz PhD); **Bedfordshire Hospitals NHS Foundation Trust, Bedford, UK** (P Dickson BSc); **The Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK** (M A Soltan MRCP, Prof D R Thickett FRCP); **Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK** (M A Soltan, Prof D R Thickett); **NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford and Public Health England, Oxford, UK** (Prof D W Eyre); **NIHR Oxford Biomedical Research Centre, Oxford, UK** (Prof D W Eyre, Prof D A Clifton); **Oxford-Suzhou Centre for Advanced Research, Suzhou, China** (Prof D A Clifton)

Correspondence to:
Dr Andrew A S Soltan, Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, UK
andrew.soltan@oncology.ox.ac.uk

Research in context

Evidence before this study

International consortia have highlighted the importance of adequate representation in health artificial intelligence (AI) datasets; however, systematic reviews have shown shortfalls in the diversity of publicly available training data. The 2013 Caldicott Information Governance Review made recommendations around best practices for health-care providers participating in data sharing, and the subsequent emergence of federated learning has been highlighted as a promising solution to enable contribution to medical AI development while retaining custody of protected health data. We searched PubMed, with no language restrictions, from database inception to Nov 1, 2022, for applications of federated learning in hospitals using the search terms "federated learning" AND ("hospital" OR "hospitals") AND ("screen" OR "screening" OR "diagnosis" OR "prognosis" OR "prognostication" OR "outcomes"). We retrieved 32 records, all published since 2020, of which five describe applications of federated learning to secondary care data, including the use of medical imaging (chest x-ray and computerised tomography) for diagnosis and prognostication in patients with COVID-19. To our knowledge, no studies to date describe the use of microcomputing or an embedded system alongside federated learning to assist in its deployment in hospitals, or demonstrate federated learning-driven COVID-19 screening using readily available, routinely collected vital signs and blood tests.

Added value of this study

To our knowledge, our study is the first to supply hospital groups with a federated learning software pipeline preconfigured on hardware as an embedded system client (which we introduce as full-stack federated learning), thereby addressing the need for in-house specialist technical expertise as an implementational barrier. We selected the commercially produced Raspberry Pi 4 model B for its low cost (£45–85; thereby enabling rapid scale-up) and its use of removable micro secure digital (microSD) cards for data storage, which are securely destroyed after participation is complete to prevent subsequent data leakage loss. We present the development and validation of COVID-19 screening models using federated learning across four hospital groups in the UK, extending our previous work. Our results show a large improvement in performance when training is federated, with a relatively greater performance increase for deep learning than for logistic regression, and robust and generalisable performance of the global model across the hospital groups in which it was evaluated.

Implications of all the available evidence

Full-stack federated learning addresses an implementational barrier to federated learning within secondary care settings and allows hospitals to participate in developing and validating AI models while retaining data within the organisation. Federated learning could be an enabling technology for deep learning, and microcomputing hardware could have a role in implementing full-stack federated learning in situations in which access to diverse training data, rather than computing power, is limiting.

Introduction

Ethical, legal, and technical considerations surround the use of patient data for medical artificial intelligence (AI) research. Risks of unintended use, misuse, and re-identification attacks,[1–3] coupled with organisational concerns of loss of control after data are transferred off-premises, could hamper efforts to improve diversity within training sets.[4]

Federated learning has emerged as a leading privacy-enhancing technology for the collaborative development of AI models without transferring data outside of participating organisations.[5,6] In classical machine learning, training takes place centrally where data are aggregated, whereas in federated learning data remain under the custody of the local organisation, and training and evaluation occur locally. Client–server federated learning is one such implementation in which weights within the model—not patient data—are transferred from client devices at each participating hospital to a centralised server after each round of local training, and aggregation takes place on the server to form a global model. After each round, the global model is recirculated to clients for updating and iteration.[7,8] In peer-to-peer federated learning, clients communicate directly to conduct aggregation without the involvement of a coordinating server.[7]

Federated learning could encourage health-care providers to participate in AI research, thereby reducing development time, improving representation, and facilitating international collaboration.[4] However, to date, real-world implementations in the hospital setting have been few in number,[9–12] with a majority of studies simulating deployment rather than conducting on-premises implementation.[13–15] Barriers to deployment include a need for specialist technical expertise at each participating site to set up and operate a federated client and ensuring adequate data sandboxing from live clinical systems. Successful deployments have used client–server federated learning for the prediction of mechanical ventilation or death in patients with COVID-19, using the NVIDIA (Santa Clara, CA, USA) Clara Platform,[6] and for automated boundary detection of glioblastoma using the OpenFL platform (Intel [Santa Clara, CA, USA] and University of Pennsylvania [Philadelphia, PA, USA]),[16] but do not fully detail the installation and set-up processes required to establish clients at each participating hospital. Pati and colleagues[16] noted that challenges include the substantial amount of

coordination needed between participating sites and the management of large volumes of communication.

An embedded system, in which a federated learning software pipeline is deployed alongside coupled dedicated hardware—introduced here as full-stack federated learning—could improve the ability of hospitals to participate in AI development by offering an easy-to-use solution that is accessible to any user proficient in information technology (IT). As a dedicated device, private data held on the embedded system can be sandboxed from other devices on the network. To our knowledge, no studies have previously investigated the use of embedded systems as federated learning clients in secondary care. Moreover, microcomputing could provide an inexpensive hardware strategy for full-stack federated learning where access to diverse multicentre data, rather than computing power, is performance-limiting. Competing deployment strategies could include the provisioning of software containers on to existing hospital computers, subject to limitations of hardware compatability and complexity within the set-up procedures, or providing external technical support, which has limited scalability and high cost.

We have previously developed, validated, and piloted an AI screening test for COVID-19 in emergency departments using techniques reliant upon data centralisation.[17,18] The CURIAL-Lab test aimed to reduce nosocomial transmission and ease operational pressures by using clinical data that are routinely collected within 1 h of a patient arriving in a hospital emergency department (vital signs, full blood count, liver function tests, urea and electrolytes, and C-reactive protein concentrations) to provide a high-confidence result-of-exclusion. The initial work showed that the CURIAL-Lab test had higher negative predictive value for excluding COVID-19 than lateral-flow tests, with results typically available sooner than for PCR testing.[17,18] Design considerations to prioritise confidentiality included asking National Health Service (NHS) trusts to de-identify data at source and using secure protocols for transfer to a trusted server at the University of Oxford (Oxford, UK) where analysis was conducted. However, de-identification processes can lead to a loss of informative predictors,[19] and could alone be insufficient to safeguard privacy in the event of a data leak.[20] Experimental studies have shown promising results for federated COVID-19 screening using medical imaging; however, these studies were limited to simulated settings and sample sizes were small.[9,21] Bai and colleagues[22] introduced a federated framework for the use of computerised tomography imaging to support COVID-19 diagnosis; however, training was conducted on data from a single hospital group (Wuhan Tongji Hospital Group, Wuhan, China) and a centralised UK data-lake maintained by NHS England (National Covid-19 Chest Imaging Database[23]).

To eliminate the need for the transfer of patient data and to address implementational barriers, we aimed to develop a user-friendly platform—introduced as full-stack federated learning—for federated training, calibration, and evaluation, and to demonstrate its use through the development of a COVID-19 screening test, known as CURIAL-Fed-Lab, across four hospital groups in the UK.

Methods

Study design

The four hospital groups that participated in the CURIAL-Lab study were included in the current study: Oxford University Hospitals NHS Foundation Trust (OUH), University Hospitals Birmingham NHS Foundation Trust (UHB), Bedfordshire Hospitals NHS Foundation Trust (BH), and Portsmouth Hospitals University NHS Trust (PUH). OUH, UHB, and PUH participated in federated training and calibration, and OUH, PUH, and BH participated in federated evaluation (figure 1).

Approval to use de-identified, routinely collected clinical and microbiology data from electronic health records for development and validation of artificial intelligence models to detect COVID-19 was granted by the NHS Health Research Authority (IRAS ID 281832).

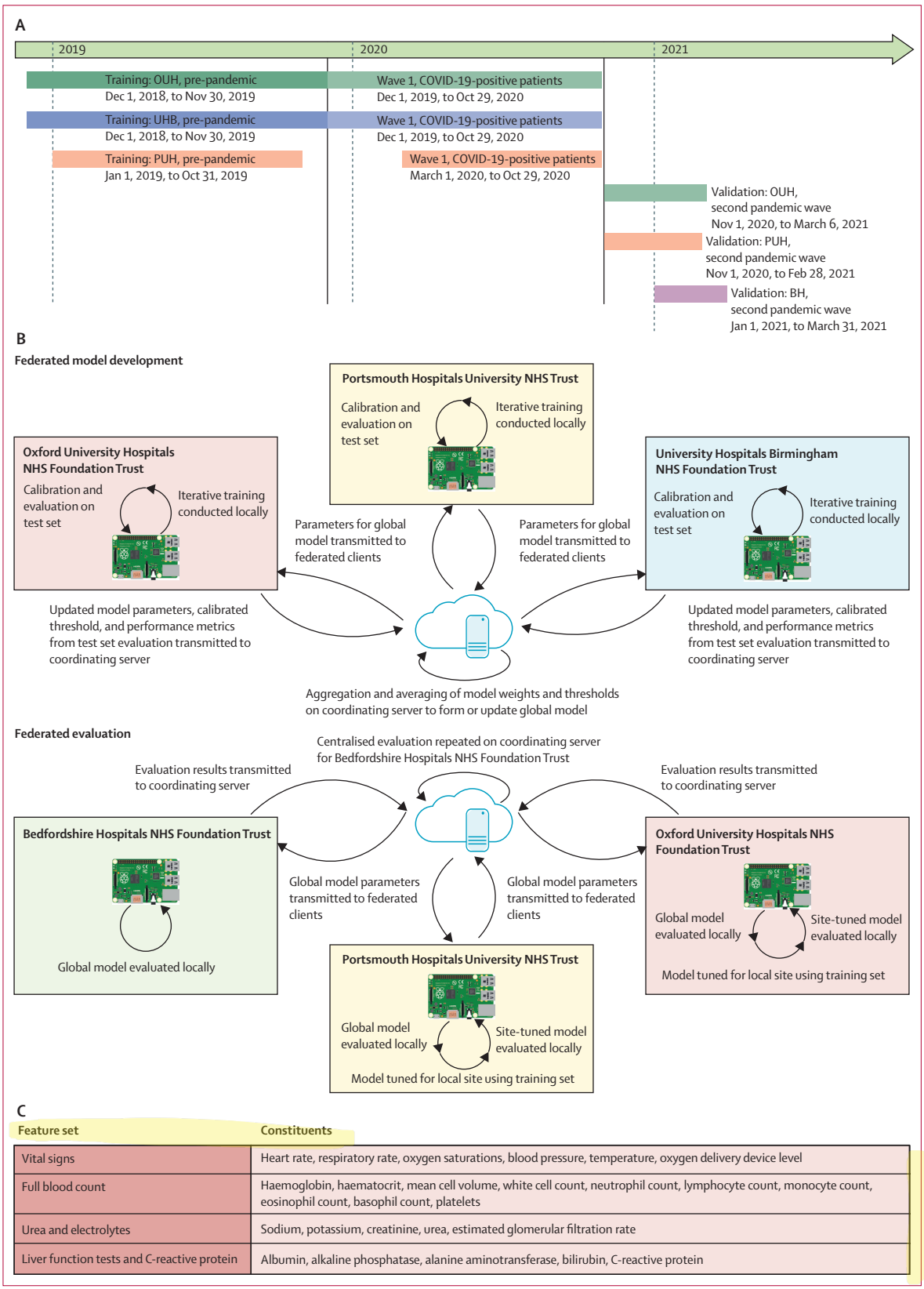
Implementation

For full-stack federated learning, we adopted a client–server architecture, supplying a Raspberry Pi 4 Model B (the client),[24] configured with at least 2 GB of random access memory and 32 GB of removable micro secure digital (microSD) storage, to participating NHS trusts or their linked research university. We preinstalled a long-term support release of Ubuntu Desktop (22.04.1 LTS), necessary dependency packages, and our custom federated learning pipeline. Clients were operated on-premises by the respective NHS trusts at PUH, UHB, and BH, and by the locally linked university at OUH (University of Oxford). The coordinating server was hosted in a dedicated virtual machine on the Microsoft Azure platform (Microsoft, Redmond, WA, USA), within an isolated virtual network. Where necessary, firewall rules were instated to permit two-way communication between client and server through a pre-agreed port. We deployed a custom analysis pipeline, preinstalled on the Raspberry Pi 4 Model B devices, to locally preprocess the de-identified data extracts, conduct data normalisation and imputation, and participate in federated training, calibration, and evaluation. On completion of participation, sites were directed to remove and securely dispose of the microSD card following local procedures.

Study populations

We provided participating NHS trusts with inclusion and exclusion criteria (appendix p 2) to enable extraction of relevant data from electronic health records, in addition to requested clinical parameters (appendix p 3). Screening against criteria, followed by de-identification

See Online for appendix



C

Feature set	Constituents
Vital signs	Heart rate, respiratory rate, oxygen saturations, blood pressure, temperature, oxygen delivery device level
Full blood count	Haemoglobin, haematocrit, mean cell volume, white cell count, neutrophil count, lymphocyte count, monocyte count, eosinophil count, basophil count, platelets
Urea and electrolytes	Sodium, potassium, creatinine, urea, estimated glomerular filtration rate
Liver function tests and C-reactive protein	Albumin, alkaline phosphatase, alanine aminotransferase, bilirubin, C-reactive protein

Figure 1: **Overview of study design**

(A) Timeline showing the derivation of training and temporally prospective evaluation cohorts. (B) Federated training and evaluation. In the model development stage, de-identified patient data are extracted by NHS trusts and loaded onto the Raspberry Pi-based federated client devices held locally within the hospital group or its linked research university. Machine learning models are trained locally and calibrated and evaluated on a locally held test set. Model weights, thresholds, and evaluation results are transmitted to a coordinating server, where aggregation and averaging is performed to form a global model. Updated weights for the new global model are transmitted to federated clients, facilitating the next round of training. 150 rounds are performed. Federated evaluation is conducted by applying the global models to temporally prospective cohorts of patients admitted to hospital during the second wave of the COVID-19 pandemic at OUH, PUH, and BH. For sites also contributing to training (OUH and PUH), an additional step of site-specific fine-tuning is conducted and the tuned model evaluated. Evaluation results are transmitted to the coordinating server for reporting. For quality assurance, centralised evaluation is also repeated on the coordinating server for BH. (C) Clinical predictors within the CURIAL-Fed-Lab model. BH=Bedfordshire Hospitals NHS Foundation Trust. NHS=National Health Service. OUH=Oxford University Hospitals NHS Foundation Trust. PUH=Portsmouth Hospitals University NHS Trust. UHB=University Hospitals Birmingham NHS Foundation Trust.

and data extraction, was done by each participating site and enforced programmatically within the analysis pipeline. Data were rendered anonymous by the clinical care team, or by informaticians employed by the extracting trusts who routinely process data as part of their NHS role, before processing in this study. Individual informed consent was not required for this study as data were routinely collected within usual care, and the study investigators did not have access to protected health information or means of re-identifying the data.

Owing to incomplete penetrance of COVID-19 testing and imperfect test sensitivity during the first wave of the pandemic, there is uncertainty in the infection status of patients presenting during this time who were untested or who tested negative. Therefore, as in our previous study,[17] for training we selected a pre-pandemic control cohort (presenting before Dec 1, 2019) to ensure the absence of disease in patients categorised as COVID-19-negative. Patients presenting during the first wave of the pandemic—defined as between Dec 1, 2019, and Oct 29, 2020—with PCR-confirmed SARS-CoV-2 infection formed the COVID-19-positive training cohort. For federated evaluation, we selected temporally prospective sets of adult patients admitted to OUH, PUH, and BH during the second pandemic wave (Nov 1, 2020, to March 31, 2021). The exact date ranges for each stage varied by site (figure 1A). Evaluation included patients receiving confirmatory molecular testing with either a positive or negative result; indeterminate or invalid results were excluded. Further information on the clinical cohorts and confirmatory testing method is provided in the appendix (pp 1–2).

Information extracted for each patient included demographics (age, sex, and ethnicity) and the initial vital signs, blood test results, blood gas measurements, and molecular SARS-CoV-2 test results that were collected on admission to hospital. Routinely conducted blood tests—full blood count, urea and electrolytes, liver function tests, and C-reactive protein concentrations—were selected because they are widely conducted within existing care pathways and results are typically available within 1 h.[17] Participating organisations were directed to load the data extracts onto the client device and activate the study application (appendix p 3).

Federated training

Feature names, result representations and units, and SARS-CoV-2 PCR results were locally preprocessed into a common data format and inclusion and exclusion criteria were programmatically enforced. Missing data were imputed by selecting the median value of the local training population, as we previously showed the stability of model performance across multiple imputation strategies.[18] Training population median values for each site were transmitted to the federated server to facilitate imputation at sites that were conducting evaluation only. As previously,[17,18] patients with PCR-confirmed SARS-CoV-2 infection during the first wave of the pandemic were matched with pre-pandemic controls across three demographic factors (ethnicity, sex, and age to within 4 years per participant). A case-to-control ratio of 1:10 was selected during training to limit the degree of class imbalance. 20% of the training set at each site, selected at random, was reserved as a test set for internal evaluation and calibration.

We conducted 150 rounds of federated training across three contributing hospital groups (OUH, PUH, and UHB), implementing the FedAvg algorithm.[8] Initial model parameters were randomly generated and clients trained a local model on their individual training sets. After local training, local models were evaluated and model parameters were transmitted by clients to the central server for aggregation and calculation of a global model. The new global model parameters were subsequently transmitted to the clients, replacing the locally held model, before the next training round. To maximise data use, we sampled each participating site (client) for every round of training. Locally held datasets were not accessible to the server during training.

We conducted federated training for two different binary classifiers aiming to predict the SARS-CoV-2 PCR result. First, as a base case, we trained a logistic regression classifier with an L2 ridge regression regularisation penalty, conducting five iterations over the training data per round. Next, we trained a deep neural network comprising an input layer, a dense hidden layer with ten nodes, a dropout regularisation layer (dropout rate 0·5) to mitigate overfitting, and an output layer. The rectified linear unit activation function was used for the hidden layers and the sigmoid activation function was used in the output layer. For updating model weights, the Adaptive Moment Estimation optimiser was used with a learning rate of 0·0001. For initial local training and for each subsequent round of federated learning, we configured the clients to iterate over the training data for up to 50 epochs with early stopping if the area under the curve on the held-out test set did not improve over 15 sequential epochs. Each client tracked the performance of its best-performing local model when evaluated on the held-out test set after each epoch, transmitting weights for this best model to the server for aggregation and for updating of the global model.

Testing and calibration

After each round of federated training, local models were calibrated by selecting the prediction threshold required to achieve a sensitivity of 85%[17,18] on the held-out test. Evaluation results for the test set, and the selected threshold, were transmitted to the coordinating server for aggregation.

Federated evaluation of the global model

We conducted federated evaluation of the global models, calibrated to 85% sensitivity, using temporally prospective

cohorts of emergency admissions to OUH, PUH, and BH during the second wave of the COVID-19 pandemic (exact date ranges varied by site; figure 1). Model predictions were evaluated by comparison with the results of confirmatory molecular testing.

For sites contributing to both federated training and federated evaluation (OUH and PUH), calibration was conducted by selecting the locally determined threshold identified during calibration on the held-out test set. Missing data were imputed using median values of the training population at the local site. For sites that were conducting federated evaluation only (BH), we selected the threshold by performing autonomous server-side averaging (mean) of the optimum local thresholds at each of the three sites contributing to training (OUH, PUH, and UHB). Missing data at BH were imputed by autonomously calculating the mean of the median population values for the three contributing sites on the evaluation server and transmitting the result to the BH client. Summary statistical measures of the results of federated evaluation were transmitted to the server for reporting.

Site-specific model tuning

We investigated the sensitivity of the global model to distribution shifts between sites as a proxy for generalisability. For sites contributing to both training and evaluation (OUH and PUH), we fine-tuned the nascent global model after each round by conducting a final training cycle on the local training set (figure 1). The performance of the fine-tuned model was assessed on the evaluation cohorts for the second pandemic wave and compared with that of the untuned global model for each round.

Centralised (server-side) evaluation

Optionally, to confirm fidelity of the federated evaluation, we repeated the evaluation of the global model for all patients admitted to BH on the coordinating server using the same threshold and imputation strategy. Summary statistical results from the federated evaluation, transmitted by the BH client, were verified to ensure they matched the results obtained from the evaluation repeated centrally. The BH data extract was transferred to the server to facilitate this verification. To understand the effects of individual features on model predictions, we calculated Shapley additive explanations values for the global models using a subset of 400 patients.[25]

Statistical analysis

Model performance was evaluated in terms of area under the receiver operating characteristic curves (AUROCs), sensitivity, specificity, positive predictive value, negative predictive value, and F_1 score. We compared the performance of locally trained models with federated global models, federated global models with site-tuned variations, and the global logistic regression model with

	Training cohorts: pre-pandemic and wave 1 COVID-19-positive patients						Evaluation cohorts: wave 2		
	Oxford University Hospitals NHS Foundation Trust		University Hospitals Birmingham NHS Foundation Trust		Portsmouth Hospitals University NHS Trust		Oxford University Hospitals NHS Foundation Trust	Portsmouth Hospitals University NHS Trust	Bedfordshire Hospitals NHS Foundation Trust
Cohort	Pre-pandemic: Dec 1, 2018, to Nov 30, 2019	Wave 1, COVID-19-positive patients: Dec 1, 2019, to Oct 29, 2020	Pre-pandemic: Dec 1, 2018, to Nov 30, 2019	Wave 1, COVID-19-positive patients: Dec 1, 2019, to Oct 29, 2020	Pre-pandemic: Jan 1, 2019, to Oct 31, 2019	Wave 1, COVID-19-positive patients: March 1, 2020, to Oct 29, 2020	Nov 1, 2020, to March 6, 2021	Nov 1, 2020, to Feb 28, 2021	Jan 1, 2021, to March 31, 2021
Patients	68 496	816	12 901	439	47 772	517	18 543	13 260	1183
Positive COVID-19 genome test	..	816 (100%)	..	439 (100%)	..	517 (100%)	1916 (10·3%)	1488 (11·2%)	145 (12·3%)
Sex									
Male	32 286 (47·1%)	435 (53·3%)	5900 (45·7%)	257 (58·5%)	20 345 (42·6%)	315 (60·9%)	9235 (49·8%)	5816 (43·9%)	629 (53·2%)
Female	36 210 (52·9%)	381 (46·7%)	7001 (54·3%)	182 (41·5%)	27 425 (57·4%)	202 (39·1%)	9308 (50·2%)	7442 (56·1%)	553 (46·8%)
Age, years	64 (44–79)	69 (54–81)	61 (40–79)	65·0 (51–81)	65 (41–79)	73 (60–83)	67 (49–80)	69 (48–82)	68 (48–82)
Ethnicity									
White	56 295 (82·2%)	554 (67·9%)	8486 (65·8%)	228 (51·9%)	37 321 (78·1%)	367 (71·0%)	14 079 (75·9%)	9954 (75·1%)	1030 (87·1%)
Not stated	8050 (11·8%)	149 (18·3%)	1231 (9·5%)	69 (15·7%)	9355 (19·6%)	131 (25·3%)	3340 (18·0%)	3014 (22·7%)	..*
South Asian	1507 (2·2%)	34 (4·2%)	1867 (14·5%)	96 (21·9%)	246 (0·5%)	..*	369 (2·0%)	62 (0·5%)	71 (6·0%)
Chinese	145 (0·2%)	..*	60 (0·5%)	..*	39 (0·1%)	..*	44 (0·2%)	14 (0·1%)	..*
Black	813 (1·2%)	28 (3·4%)	666 (5·2%)	21 (4·8%)	229 (0·5%)	..*	238 (1·3%)	72 (0·5%)	36 (3·0%)
Other	1112 (1·6%)	39 (4·8%)	347 (2·7%)	25 (5·7%)	358 (0·8%)	19 (3·7%)	347 (1·9%)	94 (0·7%)	33 (2·8%)
Mixed	574 (0·8%)	12 (1·5%)	244 (1·9%)	..*	212 (0·4%)	..*	126 (0·7%)	50 (0·4%)	13 (1·1%)

Data are n, n (%), or median (IQR). *Data from these categories were merged into the Other category in each cohort for statistical disclosure control

Table 1: **Summary population characteristics**

the global deep neural network model, within the federated pipeline, using DeLong's test.[26]

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Three NHS trusts (OUH, UHB, and PUH) participated in federated training, contributing routinely collected clinical data from 129 169 patients admitted to hospital before the pandemic and 1772 patients admitted with PCR-confirmed SARS-CoV-2 infection during the first wave of the pandemic. OUH, PUH, and BH participated in federated evaluation, comprising data from 32 986 patients admitted during the second wave of the pandemic, 3549 of whom tested positive for COVID-19 (figure 1, table 1). The prevalence of COVID-19 during the evaluation period was 11·2% at PUH, 12·3% at BH, and 10·3% at OUH, and the median ages of admitted patients were 69 years (IQR 48–82) at PUH, 68 years (48–82) at BH, and 67 years (49–80) at OUH (table 1).

To assess the effect of federation on model performance during development, we evaluated the global model on the held-out test set after each round of training (appendix p 8). Federation improved classifier stability for logistic regression, achieving optimum performance at all sites within ten rounds. The deep neural network classifier showed sustained improvement in AUROC across sequential rounds, with plateauing performance after approximately 50 rounds.

We compared the trained local models with the final federated global and site-tuned models by evaluating them on the second pandemic wave evaluation cohorts at sites participating in both training and evaluation (figure 2). Federated training significantly increased the AUROC of the logistic regression model, from 0·685 (95% CI 0·673–0·698) for the locally trained model to 0·829 (0·819–0·839) for the global model at OUH (DeLong p<0·0001) and from 0·731 (0·718–0·744) to 0·865 (0·854–0·876) at PUH (p<0·0001)—a mean increase in AUROC of 13·9% (SD 0·50%). The performance improvement due to federation was more marked for the deep neural network model: AUROC increased from 0·574 (0·560–0·589) to 0·872 (0·862–0·882) at OUH (p<0·0001), and from 0·622 (0·608–0·637) to 0·876 (0·865–0·886) at PUH (p<0·0001), a mean increase in AUROC of 27·6% (2·20%).

When the final global models were externally evaluated on a temporally prospective set of all patients admitted to BH during the second pandemic wave (Jan 1–March 31, 2021), both logistic regression (AUROC 0·878 [95% CI 0·851–0·904]) and deep neural network (0·917 [0·893–0·942]) models showed high classification performance. Federated calibration was effective, achieving sensitivities of 83·4% for the logistic regression model and 89·7% for the deep neural network model during

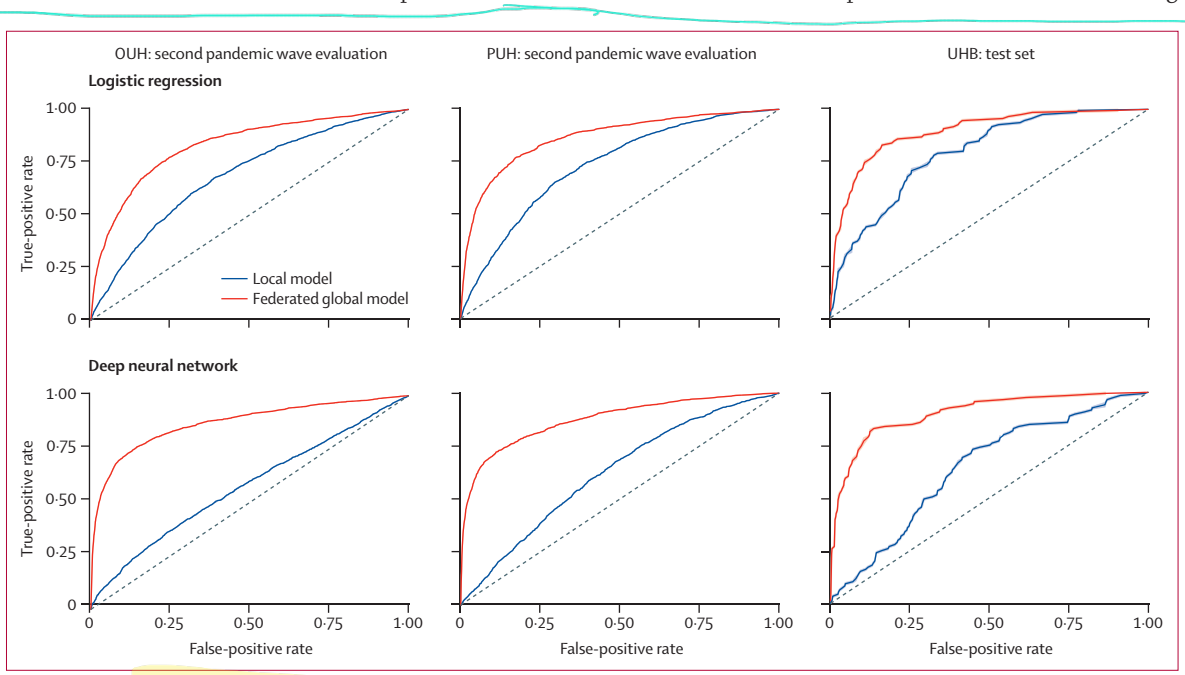


Figure 2: Comparison of locally trained and federated global models

Receiver operating characteristic curves showing the performance of locally trained models before federation (blue) and of the federated global models (orange) during evaluation for the second pandemic wave at OUH and PUH, and on the locally held test set at UHB. The area between the blue and orange curves denotes the performance improvement after 150 rounds of federation. The dashed line represents the performance of a chance predictor (AUROC of 0·5). AUROC=area under the receiver operating characteristic curve. OUH=Oxford University Hospitals NHS Foundation Trust. PUH=Portsmouth Hospitals University NHS Trust. UHB=University Hospitals Birmingham NHS Foundation Trust.

external evaluation. Both global models showed stable performance across the three evaluating sites: AUROCs ranged from 0·829 to 0·878 (95% CIs range 0·819–0·904) for the logistic regression model and from 0·872 to 0·917 (0·862–0·942) for the deep neural network model (table 2). As was observed during training, the improvement in validation performance as a result of federation was more marked for the deep neural network model (reaching a plateau after around 75–100 rounds of federation) than for the logistic regression model (plateauing after about 10 rounds; figure 3). Although the global deep neural network model outperformed the global logistic regression model at BH (DeLong p=0·0011) and at OUH (p<0·0001), the two models performed similarly at PUH (p=0.81).

Tuning of the global models for individual sites, by performing an additional round of training on the local training set before evaluation, led to a small improvement in the performance of the deep neural network model at PUH (AUROC improvement of <0·01; DeLong p=0·0014) but not at OUH (p=0·26). For the logistic regression model, site-specific fine-tuning did not improve performance (p=0·27 at PUH and p=0·63 at OUH; table 2, figure 3). This finding suggests low levels of distribution shifts in predictors between sites and high generalisability of the global models. Iteration times using the federated clients were approximately 20 s for the logistic regression model and 30 s for the deep neural network model—inclusive of training, fine-tuning, and evaluation.

Coefficient analysis of the logistic regression global model showed that granulocyte counts (neutrophils and eosinophils), albumin concentrations, and respiratory rate had the greatest effect on model predictions. This finding is consistent with the results of previous work[17,18] and with the recognised roles of these predictors in the inflammatory response. However, different from previous results, haematocrit had a relatively larger coefficient, possibly reflecting that coefficient analysis could be affected by multicolinearity. Shapley additive explanations values, which provide a quantitative measure of the effect of a feature on the predictions of a model, identified similar features as having the greatest effects on the predictions of the logistic regression global model (figure 4). For the deep neural network model, Shapley additive explanations values showed that eosinophil count has the greatest effect on model predictions.

Discussion

We trained, calibrated, and validated a screening test for COVID-19 across four hospital groups without

	AUROC	Sensitivity	Specificity	Accuracy	Positive predictive value	Negative predictive value	F_1
Oxford University Hospitals NHS Foundation Trust							
Logistic regression							
Local model	0·685 (0·673–0·698)	86·8% (85·3–88·3)	32·7% (32·0–33·4)	38·3% (37·6–39·0)	12·9% (12·4–13·5)	95·6% (95·0–96·1)	0·225
Federated global model	0·829 (0·819–0·839)	81·1% (79·3–82·8)	70·1% (69·4–70·8)	71·2% (70·6–71·9)	23·8% (22·8–24·9)	97·0% (96·7–97·3)	0·368
Federated site-tuned model	0·83 (0·819–0·84)	80·0% (78·1–81·7)	71·4% (70·7–72·1)	72·3% (71·6–72·9)	24·4% (23·3–25·5)	96·9% (96·5–97·2)	0·374
Deep neural network							
Local model	0·574 (0·56–0·589)	83·4% (81·6–85·0)	20·6% (20·0–21·2)	27·1% (26·5–27·7)	10·8% (10·3–11·3)	91·5% (90·6–92·3)	0·191
Federated global model	0·872 (0·862–0·882)	80·8% (79·0–82·5)	78·6% (78·0–79·3)	78·9% (78·3–79·5)	30·4% (29·1–31·7)	97·3% (97·0–97·5)	0·442
Federated site-tuned model	0·873 (0·863–0·883)	81·1% (79·2–82·7)	78·0% (77·3–78·6)	78·3% (77·7–78·9)	29·8% (28·5–31·0)	97·3% (97·0–97·5)	0·435
Portsmouth Hospitals University NHS Trust							
Logistic regression							
Local model	0·731 (0·718–0·744)	81·8% (79·7–83·7)	49·7% (48·8–50·6)	53·3% (52·5–54·2)	17·1% (16·2–18·0)	95·6% (95·0–96·1)	0·282
Federated global model	0·865 (0·855–0·876)	78·2% (76·1–80·2)	81·0% (80·3–81·7)	80·7% (80·0–81·3)	34·2% (32·6–35·8)	96·7% (96·3–97·0)	0·476
Federated site-tuned model	0·867 (0·856–0·878)	74·2% (71·9–76·4)	85·7% (85·0–86·3)	84·4% (83·8–85·0)	39·6% (37·8–41·4)	96·3% (96·0–96·7)	0·516
Deep neural network							
Local model	0·622 (0·608–0·637)	74·5% (72·3–76·7)	43·8% (42·9–44·7)	47·3% (46·4–48·1)	14·4% (13·6–15·2)	93·2% (92·5–93·8)	0·241
Federated global model	0·876 (0·865–0·886)	77·2% (74·9–79·2)	82·3% (81·6–82·9)	81·7% (81·0–82·3)	35·5% (33·8–37·1)	96·6% (96·2–96·9)	0·486
Federated site-tuned model	0·883 (0·873–0·893)	78·2% (76·1–80·2)	82·7% (82·0–83·4)	82·2% (81·6–82·9)	36·4% (34·7–38·1)	96·8% (96·4–97·1)	0·497
Bedfordshire Hospitals NHS Foundation Trust							
Logistic regression: federated global model	0·878 (0·851–0·904)	83·4% (76·6–88·6)	73·6% (70·8–76·2)	74·8% (72·3–77·2)	30·6% (26·3–35·3)	97·0% (95·5–97·9)	0·448
Deep neural network: federated global model	0·917 (0·893–0·942)	89·7% (83·6–93·6)	76·6% (73·9–79·1)	78·2% (75·7–80·5)	34·9% (30·2–39·8)	98·1% (97·0–98·9)	0·502

Data are measure (95% CI) or measure. Calibration was performed locally during training for sites participating in both federated training and evaluation (Oxford University Hospitals NHS Foundation Trust and Portsmouth Hospitals University NHS Trust), and was federated for sites participating only in evaluation (Bedfordshire Hospitals NHS Foundation Trust). Performance of the federated models is after 150 rounds of federated training. AUROC=area under the receiver operating characteristic curve.

Table 2: **Performance of calibrated local and federated models in identifying patients being admitted to hospital with COVID-19 when evaluated for patients admitted during the second pandemic wave**

centralising patient data, developing a user-friendly embedded system (full-stack federated learning) to enable participation in the study without specialist technical expertise. To our knowledge, our study is the first to use microcomputing within an embedded system

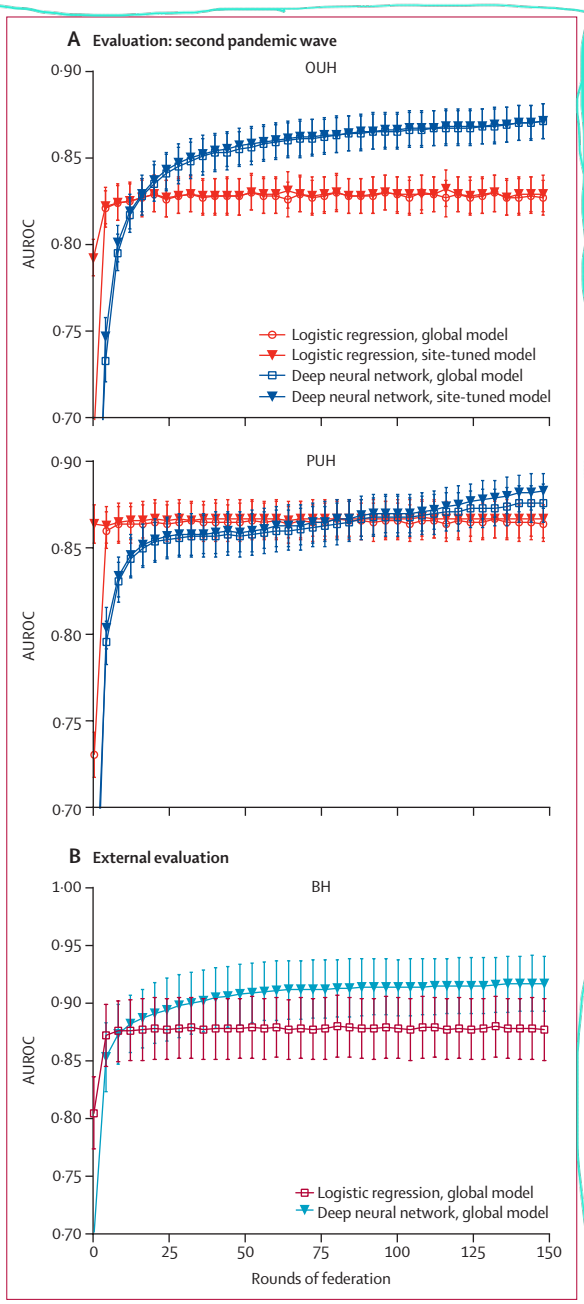


Figure 3: **Effect of federated training on the performance of logistic regression and deep neural network models**
(A) Evaluation of the global and site-tuned models on evaluation sets of patients admitted to OUH and PUH during the second pandemic wave. (B) External evaluation of the global model for patients admitted to BH during the second wave of the pandemic. Data are AUROC (95% CI). AUROC=area under the receiver operating characteristic curve. BH=Bedfordshire Hospitals NHS Foundation Trust. OUH=Oxford University Hospitals NHS Foundation Trust. PUH=Portsmouth Hospitals University NHS Trust.

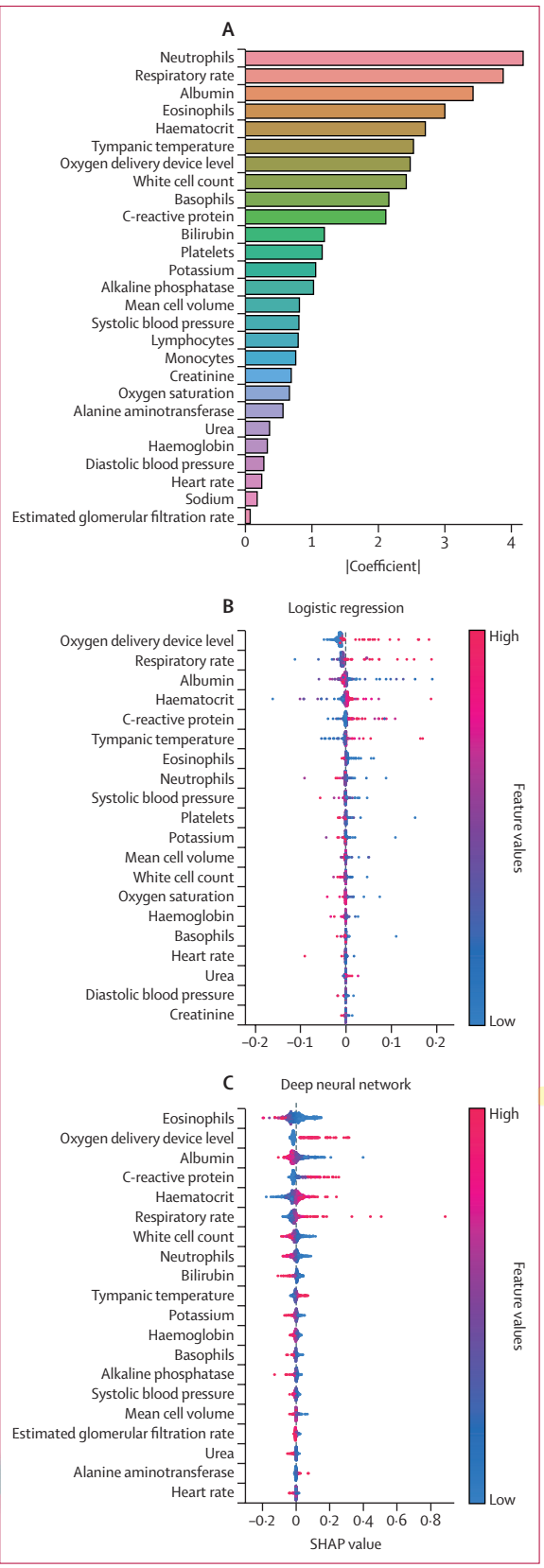


Figure 4: **Explainability analyses**
(A) Logistic regression coefficient scalars within the final global model. (B) SHAP values for the 20 features with the greatest effect on predictions made by the logistic regression global model. (C) As (B) but for the deep neural network model. SHAP values were calculated during centralised external validation for BH and shown as beeswarm plots. Each dot represents a patient attending BH during the evaluation period. Positive SHAP values indicate a change in the expected model prediction towards testing positive for COVID-19. Features are shown in descending order of mean absolute SHAP value, with the most impactful features shown at the top. BH=Bedfordshire Hospitals NHS Foundation Trust. SHAP=Shapley additive explanations.

to aid the deployment of federated learning in the hospital setting. Our microcomputing implementation uses commercially available hardware that can be rapidly scaled at low per-site cost (£45–85). We propose that federated learning has the potential to become a new standard-of-practice for privacy-preserving health data research, reducing barriers to participation and bias within training sets[5] and enabling cross-border collaboration while upholding data sovereignty.

Our results show that federation significantly improved performance compared with training on data from a single site (figure 2), bringing our models' performance into a clinically acceptable range. We observed a more marked increase in performance for deep neural networks than for logistic regression models, in keeping with the findings from other applications of federated deep learning.[27] This result is possibly reflective of deep neural networks requiring large quantities of data to extract high-level features, indicating that federated learning could be an enabling technology for deep learning in health AI.[27] When compared with the original CURIAL-Lab model, which was based on the XGBoost algorithm, this global deep neural network model achieves higher performance in a comparable evaluation at BH (CURIAL-Lab model AUROC 0·881 [95% CI 0·851–0·912][18] *vs* CURIAL-Fed-Lab model 0·917 [0·893–0·942]). In this work, distribution shifts between sites were small and global model generalisability was high; however, this could vary in different clinical scenarios in which predictors show greater inter-site variation, for example where there are differences in acquisition method or sample preparation.

Federated learning is considered private by design because data remain with the health-care provider; however, additional considerations are needed within the federation to address risks of unauthorised on-device access, malicious code injection, and information leakage. The use of single-purpose client and server hardware reduced the risk of inadvertent Trojan attack or sandbox violation. We selected the most recent long-term support release of Ubuntu Desktop (22.04.1 LTS), a free, open, and commercially supported Linux distribution, as it provides a graphical user interface for ease of use and is supported by software dependencies for our pipeline. Clients were secured in line with local requirements, and participating sites were asked to physically safeguard devices following local processes for IT hardware holding pseudonymised data. Data were held on the client in pseudonymised form for the period of analysis only, protected by the site network's firewall, and clients were switched off when not in use. Sites were asked to remove and destroy the microSD storage disk on completion of participation. Where required, firewall rules were instated by local IT or network security teams to allow two-way traffic communication between the device and the coordinating server via a single, pre-agreed port. The coordinating server was subject to the security considerations of the Azure platform.[28] External communication was restricted to the pre-agreed port only, and the server was switched off when not in use for the present study. Messages between client and server contained only weights from within the trained model or summary results of evaluation, providing inherent protection against leakage if intercepted. Moreover, as the model architectures contained many fewer parameters than there were training data examples at each site, the risk of raw training data being memorised within model updates was low. Future work could investigate the use of federated policies and authorisation controls.[29]

Notable limitations of our study included that previous knowledge of the data format was required to allow harmonisation of feature names, unit values, and the representation of out-of-bounds values between sites. We approached this by providing trusts with a data specification and dictionary; however, future work could explore a role for complimentary privacy-enhancing technologies, such as differential privacy or synthetic data, where a more in-depth knowledge of the dataset is required.[30,31] Further, future work could seek to implement a fully autonomous data extraction and harmonisation pathway through direct integration with the electronic health record, alongside appropriate governance infrastructure to ensure that the ethical considerations of new clinical aims are comprehensively evaluated before deployment.[32,33] Direct integration to the electronic health record could be supported by specifying features using standardised notation, such as Logical Observation Identifier Names and Codes. The distributed nature of federated learning required that sensitivity and subgroup analyses are defined a priori, because only model weights and evaluation results are transmitted, potentially limiting the ability of researchers to investigate trends discovered within early results. Future work could investigate secure methods to enable post-hoc sensitivity analyses and updating of the code while protecting against code injection, in addition to novel efficient strategies for hyperparameter tuning.[34] Federated learning, in combination with other privacy-enhancing technologies, does not create a trustless system and continues to require professional conduct during the manual stages of data and device handling. Finally, the small size of microcomputing hardware could increase its susceptibility to loss or theft, requiring greater consideration of physical security measures than for larger devices.

Because federated learning is within its infancy in health care, challenges to the collaboration included identifying local technical and governance stakeholders and demonstrating the rationale for federation. We anticipate that onboarding could become easier as awareness of federated learning increases. We made the source code available for review, and at one site (PUH) provided security teams with example contents of

messages transmitted between clients and server to demonstrate abstraction from the raw data. Because we were deploying an embedded system, technical discussions did not need to cover existing NHS trust hardware and focused around establishing access controls to the system. Once in place, instructions to transfer the electronic health record extracts onto the system and connect to the aggregation server were provided by email, guided by numbered desktop shortcuts. We experienced connection dropping during the federation, requiring the process to be restarted, which could be addressed as the software matures. Regulatory precedent for the live deployment of models that can update as more data become available is currently limited by the need for performance assurance. However, the US Food and Drug Administration has begun consultation on new proposals inclusive of support for continuous learning algorithms, enabling the benefits of federated learning to be maximally realised.[35]

We use the Raspberry Pi 4 Model B device owing to commercial availability, high levels of support, and inexpensive removable storage medium (microSD; <£9 for 32 GB). To aid rapid deployment at scale, the microSD card of a configured Raspberry Pi can be imaged and cloned for onboarding new sites. Further, as the microSD cards are interchangeable, new clinical indications can be addressed by dispatching updated microSD cards to participating hospitals on receipt of appropriate approvals, maintaining the benefits of ease of use by enabling configuration to take place centrally before dispatch. Iteration times were approximately 20 s for the logistic regression model and 30 s for the deep neural network model, supporting the use of microcomputing for this tabular learning task. Although implementing full-stack federated learning using a microcomputer is effective for cases in which access to diverse training data rather than computational power is performance-limiting, hardware with greater computational power might be required for more intensive tasks. Because our pipeline is designed for the well supported Ubuntu operating system, the software can be adapted to more powerful hardware where appropriate for computer vision and natural language applications. Moreover, where local expertise is available, the federated learning pipeline can be deployed on hospital-owned hardware.

In conclusion, we present an inexpensive and easy-to-use embedded system for federated learning and successfully deploy this system in the real-world secondary care setting. Future work could evaluate the effects on model fairness due to the improved representation in training data brought about by federation, and applications to new clinical questions.[27]

Contributors

AASS conceived of and designed the study with input from TZ, AT, and DAC. AASS wrote the federated learning client code, conducted the client and server implementations, conducted the analyses, and wrote the manuscript. AT provided support with development of the federated learning set-up. JY supported with earlier versions of code to harmonise data at two of the sites. DWE performed data extraction at OUH. PD (BH), LGD'C and AC (PUH), AASS (OUH), and MAS and DRT (UHB) operated the client devices at the respective sites. Although, by design, patient data were not transferred within the federation, AASS, JY, DWE, and DAC have previously had access to and verified the raw data within a previous related evaluation study (Soltan et al[18]). PD accessed and verified the data at BH, MAS and DRT accessed and verified the data at UHB, and LGD'C accessed and verified the data at PUH. All authors reviewed and edited the manuscript. All authors had final responsibility for the decision to submit for publication.

Declaration of interests

DAC reports grant funding from the Wellcome Trust (217650/Z/19/Z), Research Councils UK (EP/V003321/1, EP/W031744/1, EP/P009824/1, EP/N024966/1, EP/N027000/1, and EP/N020774/1), and GlaxoSmithKline; personal fees from Oxford University Innovation, Sensyne Health, and Bristol Myers Squibb; and has received honoraria for patents held by Oxford University Innovation, all outside the submitted work. AC reports grants from Exhalation Technology, Asthma UK (AUK-IG-2016-357), the National Institute for Health and Care Research (NIHR; II-LA-1117-20002 and AI_AWARD02031), Glyconics, Boehringer Ingelheim (2020/M/015), and Sanofi, all outside the submitted work. DWE reports personal fees from Gilead Sciences, outside the submitted work. All other authors declare no competing interests.

Data sharing

The code for the federated learning pipeline and analyses is available online via GitHub (https://github.com/andrewsoltan/Curial-Federated-Learning-Manuscript). Data from OUH are available from the Infections in Oxfordshire Research Database (https://oxfordbrc.nihr.ac.uk/research-themes/modernising-medical-microbiology-and-big-infection-diagnostics/infections-in-oxfordshire-research-database-iord/), subject to an application that meets the ethical and governance requirements of the database. Data from UHB, PUH, and BH are available on request to the respective trusts subject to NHS Health Research Authority requirements.

Acknowledgments

We express our sincere thanks to all patients and staff across the four participating NHS trusts: Oxford University Hospitals NHS Foundation Trust, University Hospitals Birmingham NHS Foundation Trust, Bedfordshire Hospitals NHS Foundation Trust, and Portsmouth Hospitals University NHS Trust. This study was supported by the Wellcome Trust/University of Oxford Medical and Life Sciences Translational Fund (award 0009350) and the NIHR Oxford Biomedical Research Centre. AASS is an NIHR Academic Clinical Fellow (award ACF-2020-13-015). DWE is a Robertson Foundation Fellow and an NIHR Oxford Biomedical Research Centre Senior Fellow. DAC was supported by an RAEng Research Chair, NIHR Research Professorship, the InnoHK Hong Kong Centre for Cerebro-cardiovascular Health Engineering, the Oxford Pandemic Sciences Institute, and the Oxford-Suzhou Centre for Advanced Research (Suzhou, China). JY is a Marie Skłodowska-Curie Fellow, under the EU's Horizon 2020 research and innovation programme (grant agreement 955681, MOIRA). The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or the Wellcome Trust.

References

1 Price WN 2nd, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019; **25:** 37–43.
2 Department of Health and Social Care. Better, broader, safer: using health data for research and analysis. April 7, 2022. https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis (accessed Jan 3, 2023).
3 Henriksen-Bulmer J, Jeary S. Re-identification attacks—a systematic literature review. *Int J Inf Manage* 2016; **36:** 1184–92.
4 Ganapathi S, Palmer J, Alderman JE, et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat Med* 2022; **28:** 2232–33.

5 Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *NPJ Digit Med* 2020; **3:** 119.

6 Dayan I, Roth HR, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nat Med* 2021; **27:** 1735–43.

7 Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: strategies for improving communication efficiency. NIPS Workshop on Private Multi-Party Machine Learning; Dec 9, 2016.

8 McMahan HB, Moore E, Ramage D, Hampson S, Agüera y Arcas B. Communication-efficient learning of deep networks from decentralized data. *PMLR* 2017; **54:** 1273–82.

9 Naz S, Phan KT, Chen YPP. A comprehensive review of federated learning for COVID-19 detection. *Int J Intell Syst* 2022; **37:** 2371–92.

10 Crowson MG, Moukheiber D, Arévalo AR, et al. A systematic review of federated learning applications for biomedical data. *PLOS Digit Health* 2022; **1:** e0000033.

11 Rajendran S, Obeid JS, Binol H, et al. Cloud-based federated learning implementation across medical centers. *JCO Clin Cancer Inform* 2021; **5:** 1–11.

12 Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. *Int J Med Inform* 2018; **112:** 59–67.

13 Durga R, Poovammal E. FLED-Block: federated learning ensembled deep learning blockchain model for COVID-19 prediction. *Front Public Health* 2022; **10:** 892499.

14 Liang H, Guo Y, Chen X, et al. Artificial intelligence for stepwise diagnosis and monitoring of COVID-19. *Eur Radiol* 2022; **32:** 2235–45.

15 Chang K, Balachandar N, Lam C, et al. Distributed deep learning networks among institutions for medical imaging. *J Am Med Inform Assoc* 2018; **25:** 945–54.

16 Pati S, Baid U, Edwards B, et al. Federated learning enables big data for rare cancer boundary detection. *Nat Commun* 2022; **13:** 7346.

17 Soltan AAS, Kouchaki S, Zhu T, et al. Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test. *Lancet Digit Health* 2021; **3:** e78–87.

18 Soltan AAS, Yang J, Pattanshetty R, et al. Real-world evaluation of rapid and laboratory-free COVID-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening. *Lancet Digit Health* 2022; **4:** e266–78.

19 Carvalho T, Moniz N, Faria P, Antunes L. Towards a data privacy-predictive performance trade-off. *Expert Syst Appl* 2023; **223:** 119785.

20 Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* 2019; **10:** 3069.

21 Yan B, Wang J, Cheng J, et al. Experiments of federated learning for COVID-19 chest x-ray images. In: Sun X, Zhang X, Xia Z, Bertino E, eds. Advances in artificial intelligence and security. 7th International Conference, ICAIS 2021; July 19–23, 2021. Cham: Springer Cham, 2021: 41–53.

22 Bai X, Wang H, Ma L, et al. Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence. *Nat Mach Intell* 2021; **3:** 1081–89.

23 Cushnan D, Bennett O, Berka R, et al. An overview of the National COVID-19 Chest Imaging Database: data quality and cohort analysis. *Gigascience* 2021; **10:** giab076.

24 Raspberry Pi (Trading). Raspberry Pi 4 Model B Datasheet. June, 2019. https://datasheets.raspberrypi.com/rpi4/raspberry-pi-4-datasheet.pdf (accessed Jan 3, 2023).

25 Lundberg S, Lee S-I. A unified approach to interpreting model predictions. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; December, 2017, 4768–77.

26 Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett* 2014; **21:** 1389–93.

27 Kairouz P, McMahan HB, Avent B, et al. Advances and open problems in federated learning. *Found Trends Mach Learn* 2019; **14:** 1–210.

28 Microsoft Azure. Introduction to Azure security. https://learn.microsoft.com/en-us/azure/security/fundamentals/overview (accessed Dec 5, 2023).

29 NVIDIA. Identity security. https://nvflare.readthedocs.io/en/main/user_guide/security/identity_security.html#federated-authorization (accessed Dec 5, 2023).

30 Dwork C, Roth A. The algorithmic foundations of differential privacy. *Found Trends Theor Comput Sci* 2014; **9:** 211–407.

31 Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng* 2021; **5:** 493–97.

32 Sheikh A, Anderson M, Albala S, et al. Health information technology and digital innovation for national learning health and care systems. *Lancet Digit Health* 2021; **3:** e383–96.

33 Mandl KD, Gottlieb D, Mandel JC, et al. Push button population health: the SMART/HL7 FHIR Bulk Data Access application programming interface. *NPJ Digit Med* 2020; **3:** 151.

34 Guo P, Yang D, Hatamizadeh A, et al. Auto-FedRL: federated hyperparameter optimization for multi-institutional medical image segmentation. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, eds. Computer vision – ECCV 2022: 17th European Conference; Oct 23–27, 2022. Cham: Springer Cham, 2022: 437–55.

35 Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. *Lancet Digit Health* 2021; **3:** e337–38.