nature medicine

ARTICLES

https://doi.org/10.1038/s41591-021-01506-3

Federated learning for predicting clinical outcomes in patients with COVID-19

Ittai Dayan[1,56], Holger R. Roth[2,56], Aoxiao Zhong[3,4,56], Ahmed Harouni[2], Amilcare Gentili[5], Anas Z. Abidin[2], Andrew Liu[2], Anthony Beardsworth Costa[6], Bradford J. Wood[7,8], Chien-Sung Tsai[9], Chih-Hung Wang[10,11], Chun-Nan Hsu[12], C. K. Lee[2], Peiying Ruan[2], Daguang Xu[2], Dufan Wu[3], Eddie Huang[2], Felipe Campos Kitamura[13], Griffin Lacey[2], Gustavo César de Antônio Corradi[13], Gustavo Nino[14], Hao-Hsin Shin[15], Hirofumi Obinata[16], Hui Ren[3], Jason C. Crane[17], Jesse Tetreault[2], Jiahui Guan[2], John W. Garrett[18], Joshua D. Kaggie[19], Jung Gil Park[20], Keith Dreyer[1,21], Krishna Juluru[15], Kristopher Kersten[2], Marcio Aloisio Bezerra Cavalcanti Rockenbach[21], Marius George Linguraru[22,23], Masoom A. Haider[24,25], Meena AbdelMaseeh[25], Nicola Rieke[2], Pablo F. Damasceno[17], Pedro Mario Cruz e Silva[2], Pochuan Wang[26,27], Sheng Xu[7,8], Shuichi Kawano[16], Sira Sriswasdi[28,29], Soo Young Park[30], Thomas M. Grist[31], Varun Buch[21], Watsamon Jantarabenjakul[32,33], Weichung Wang[26,27], Won Young Tak[30], Xiang Li[3], Xihong Lin[34], Young Joon Kwon[6], Abood Quraini[2], Andrew Feng[2], Andrew N. Priest[35], Baris Turkbey[8,36], Benjamin Glicksberg[37], Bernardo Bizzo[21], Byung Seok Kim[38], Carlos Tor-Díez[22], Chia-Cheng Lee[39], Chia-Jung Hsu[39], Chin Lin[40,41,42], Chiu-Ling Lai[43], Christopher P. Hess[17], Colin Compas[2], Deepeksha Bhatia[2], Eric K. Oermann[44], Evan Leibovitz[21], Hisashi Sasaki[16], Hitoshi Mori[16], Isaac Yang[2], Jae Ho Sohn[17], Krishna Nand Keshava Murthy[15], Li-Chen Fu[45], Matheus Ribeiro Furtado de Mendonça[13], Mike Fralick[46], Min Kyu Kang[20], Mohammad Adil[2], Natalie Gangai[15], Peerapon Vateekul[47], Pierre Elnajjar[15], Sarah Hickman[19], Sharmila Majumdar[17], Shelley L. McLeod[48,49], Sheridan Reed[7,8], Stefan Gräf[50], Stephanie Harmon[8,51], Tatsuya Kodama[16], Thanyawee Puthanakit[32,33], Tony Mazzulli[52,53,54], Vitor Lima de Lavor[13], Yothin Rakvongthai[55], Yu Rim Lee[30], Yuhong Wen[2], Fiona J. Gilbert[19,56], Mona G. Flores[2,56] ✉ and Quanzheng Li[3,56]

Federated learning (FL) is a method used for training artificial intelligence models with data from multiple sources while maintaining data anonymity, thus removing many barriers to data sharing. Here we used data from 20 institutes across the globe to train a FL model, called EXAM (electronic medical record (EMR) chest X-ray AI model), that predicts the future oxygen requirements of symptomatic patients with COVID-19 using inputs of vital signs, laboratory data and chest X-rays. EXAM achieved an average area under the curve (AUC) >0.92 for predicting outcomes at 24 and 72 h from the time of initial presentation to the emergency room, and it provided 16% improvement in average AUC measured across all participating sites and an average increase in generalizability of 38% when compared with models trained at a single site using that site's data. For prediction of mechanical ventilation treatment or death at 24 h at the largest independent test site, EXAM achieved a sensitivity of 0.950 and specificity of 0.882. In this study, FL facilitated rapid data science collaboration without data exchange and generated a model that generalized across heterogeneous, unharmonized datasets for prediction of clinical outcomes in patients with COVID-19, setting the stage for the broader use of FL in healthcare.

The scientific, academic, medical and data science communities have come together in the face of the COVID-19 pandemic crisis to rapidly assess novel paradigms in artificial intelligence (AI) that are rapid and secure, and potentially incentivize data sharing and model training and testing without the usual privacy and data ownership hurdles of conventional collaborations[1,2]. Healthcare providers, researchers and industry have pivoted their focus to address unmet and critical clinical needs created by the crisis, with remarkable results[3–9]. Clinical trial recruitment has been expedited and facilitated by national regulatory bodies and an international cooperative spirit[10–12]. The data analytics and AI disciplines have always fostered open

A full list of affiliations appears at the end of the paper.

Table 1 | EMR data used in the EXAM study

Category	Subcategory	Component name	Definition	Units	LOINC code
Demographic	–	Patient age	–	Years	30525-0
Imaging	Portable CXR	–	AP or PA portable CXR	–	36554-4
Lab value	C-reactive protein	C-reactive protein	Blood c-reactive protein concentration	mg l^{-1}	1988-5
Lab value	Complete blood count (CBC)	Neutrophils	Blood absolute neutrophils	10^9 l^{-1}	751-8
Lab value	CBC	White blood cells	Blood white blood cell count	10^9 l^{-1}	33256-9
Lab value	D-dimer	D-dimer	Blood D-dimer concentration	ng ml^{-1}	7799-0
Lab value	Lactate	Lactate	Blood lactate concentration	mmol l^{-1}	2524-7
Lab value	Lactate dehydrogenase	LDH	Blood LDH concentration	U l^{-1}	2532-0
Lab value	Metabolic panel	Creatinine	Blood creatinine concentration	mg dl^{-1}	2160-0
Lab value	Procalcitonin	Procalcitonin	Blood procalcitonin concentration	ng ml^{-1}	33959-8
Lab value	Metabolic panel	eGFR	Estimated glomerular filtration rate	ml min^{-1} 1.73 m^{-2}	69405-9
Lab value	Troponin	Troponin-T	Blood troponin concentration	ng ml^{-1}	67151-1
Lab value	Hepatic panel	AST	Blood aspartate aminotransferase concentration	IU l^{-1}	1920-8
Lab value	Metabolic panel	Glucose	Blood glucose concentration	mg dl^{-1}	2345-7
Vital sign	–	Oxygen saturation	Oxygen saturation	%	59408-5
Vital sign	–	Systolic blood pressure	Systolic BP	mmHg	8480-6
Vital sign	–	Diastolic blood pressure	Diastolic BP	mmHg	8462-4
Vital sign	–	Respiratory rate	Respiratory rate	Breaths min^{-1}	9279-1
Vital sign		COVID PCR test	PCR for RNA (not used as input to model)		95425-5
Vital sign	Oxygen device used in ED	Oxygen device	Ventilation, high-flow/NIV, low-flow, room air	–	41925-9
Outcome	24-h oxygen device	Oxygen device	Ventilation, high-flow/NIV, low-flow, room air	–	41925-9
Outcome	72-h oxygen device	Oxygen device	Ventilation, high-flow/NIV, low-flow, room air	–	41925-9
Outcome	Death	–	–	–	–
Outcome	Time of death	–	–	Hours	–

and collaborative approaches, embracing concepts such as open-source software, reproducible research, data repositories and making available anonymized datasets publicly[13,14]. The pandemic has emphasized the need to expeditiously conduct data collaborations that empower the clinical and scientific communities when responding to rapidly evolving and widespread global challenges. Data sharing has ethical, regulatory and legal complexities that are underscored, and perhaps somewhat complicated, by the recent entrance of large technology companies into the healthcare data world[15–17].

A concrete example of these types of collaboration is our previous work on an AI-based SARS-COV-2 clinical decision support (CDS) model. This CDS model was developed at Mass General Brigham (MGB) and was validated across multiple health systems' data. The inputs to the CDS model were chest X-ray (CXR) images, vital signs, demographic data and laboratory values that were shown in previous publications to be predictive of outcomes of patients with COVID-19[18–21]. CXR was selected as the imaging input because it is widely available and commonly indicated by guidelines such as those provided by ACR[22], the Fleischner Society[23], the WHO[24], national thoracic societies[25], national health ministry COVID handbooks and radiology societies across the world[26]. The output of the CDS model was a score, termed CORISK[27], that corresponds to oxygen support requirements and that could aid in triaging patients by frontline clinicians[28–30]. Healthcare providers have been known to prefer models that were validated on their own data[27]. To date most AI models, including the aforementioned CDS model, have been trained and validated on 'narrow' data that often lack diversity[31,32], potentially resulting in overfitting and lower generalizability. This can be mitigated by training with diverse data from multiple sites without centralization of data[33] using methods such as transfer learning[34,35] or FL. FL is a method used to train AI models on disparate data sources, without the data being transported or exposed outside their original location. While applicable to many industries, FL has recently been proposed for cross-institutional healthcare research[36].

Federated learning supports the rapid launch of centrally orchestrated experiments with improved traceability of data and assessment of algorithmic changes and impact[37]. One approach to FL, called client-server, sends an 'untrained' model to other servers ('nodes') that conduct partial training tasks, in turn sending the results back to be merged in the central ('federated') server. This is conducted as an iterative process until training is complete[36].

Governance of data for FL is maintained locally, alleviating privacy concerns, with only model weights or gradients communicated between client sites and the federated server[38,39]. FL has already shown promise in recent medical imaging applications[40–43], including in COVID-19 analysis[8,44,45]. A notable example is a

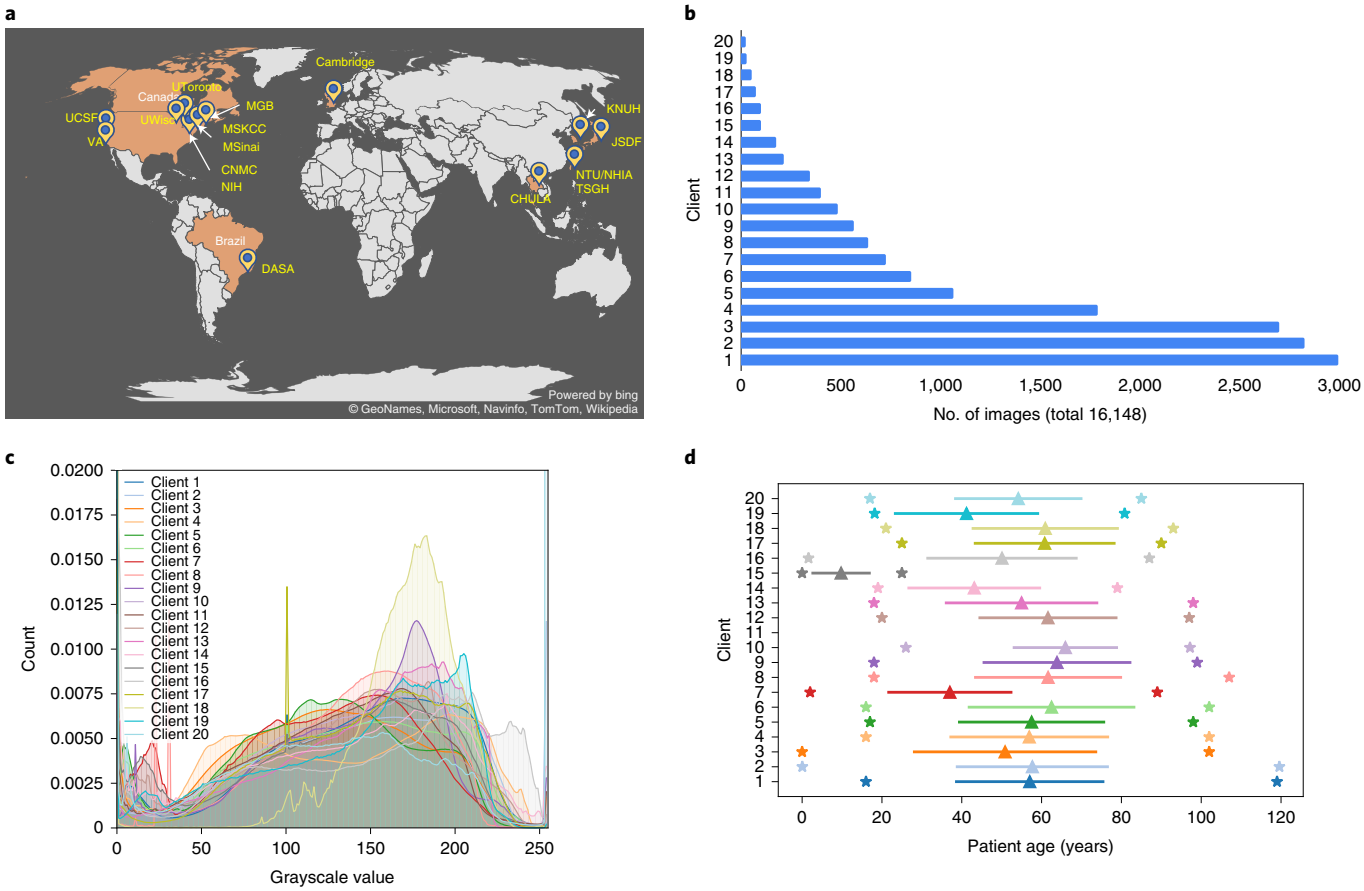


Fig. 1 | Data used in the EXAM FL study. a, World map indicating the 20 different client sites contributing to the EXAM study. **b**, Number of cases contributed by each institution or site (client 1 represents the site contributing the largest number of cases). **c**, Chest X-ray intensity distribution at each client site. **d**, Age of patients at each client site, showing minimum and maximum ages (asterisks), mean age (triangles) and standard deviation (horizontal bars). The number of samples of each client site is shown in Supplementary Table 1.

mortality prediction model in patients infected with SARS-COV-2 that uses clinical features, albeit limited in terms of number of modalities and scale[46].

Our objective was to develop a robust, generalizable model that could assist in triaging patients. We theorized that the CDS model can be federated successfully, given its use of data inputs that are relatively common in clinical practice and that do not rely heavily on operator-dependent assessments of patient condition (such as clinical impressions or reported symptoms). Rather, laboratory results, vital signs, an imaging study and a commonly captured demographic (that is, age), were used. We therefore retrained the CDS model with diverse data using a client-server FL approach to develop a new global FL model, which was named EXAM, using CXR and EMR features as input. By leveraging FL, the participating institutions would not have to transfer data to a central repository, but rather leverage a distributed data framework.

Our hypothesis was that EXAM would perform better than local models and would generalize better across healthcare systems.

Results

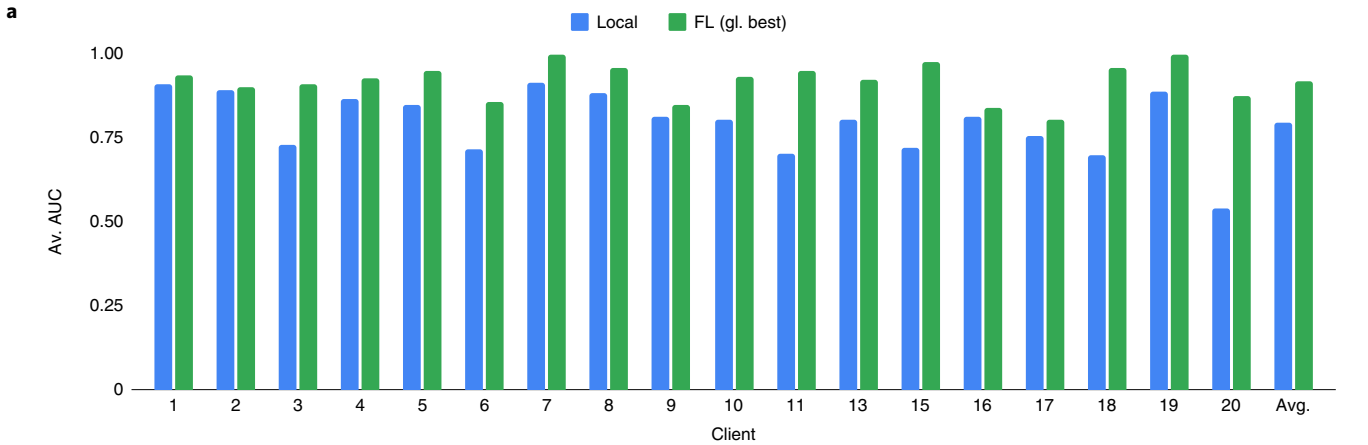
The EXAM model architecture. The EXAM model is based on the CDS model mentioned above[27]. In total, 20 features (19 from the EMR and one CXR) were used as input to the model. The outcome (that is, 'ground truth') labels were assigned based on patient oxygen therapy after 24- and 72-hour periods from initial admission to the emergency department (ED). A detailed list of the requested features and outcomes can be seen in Table 1.

The outcome labels of patients were set to 0, 0.25, 0.50 and 0.75 depending on the most intensive oxygen therapy the patient received in the prediction window. The oxygen therapy categories were, respectively, room air (RA), low-flow oxygen (LFO), high-flow oxygen (HFO)/noninvasive ventilation (NIV) or mechanical ventilation (MV). If the patient died within the prediction window, the outcome label was set to 1. This resulted in each case being assigned two labels in the range 0–1, corresponding to each of the prediction windows (that is, 24 and 72 h).

For EMR features, only the first values captured in the ED were used and data preprocessing included deidentification, missing value imputation and normalization to zero-mean and unit variance. For CXR images, only the first obtained in the ED was used.

The model therefore fuses information from both EMR and CXR features, using a 34-layer convolutional neural network (ResNet34) to extract features from a CXR and a Deep & Cross network to concatenate the features together with the EMR features (for more expanded details, see Methods). The model output is a risk score, termed the EXAM score, which is a continuous value in the range 0–1 for each of the 24- and 72-hour predictions corresponding to the labels described above.

Federating the model. The EXAM model was trained using a cohort of 16,148 cases, making it not only among the first FL models for COVID-19 but also a very large and multicontinent development project in clinically relevant AI (Fig. 1a,b). Data between sites were not harmonized before extraction and, in light of real-life



Client	1	2	3	4	5	6	7	8	9	10	11	13	15	16	17	18	19	20	Av.
Local	0.910	0.892	0.731	0.869	0.848	0.716	0.916	0.887	0.816	0.803	0.702	0.805	0.722	0.812	0.755	0.698	0.889	0.542	0.795
FL (gl. best)	0.938	0.902	0.912	0.929	0.950	0.857	1.000	0.961	0.849	0.935	0.950	0.925	0.979	0.839	0.806	0.958	1.000	0.875	0.920

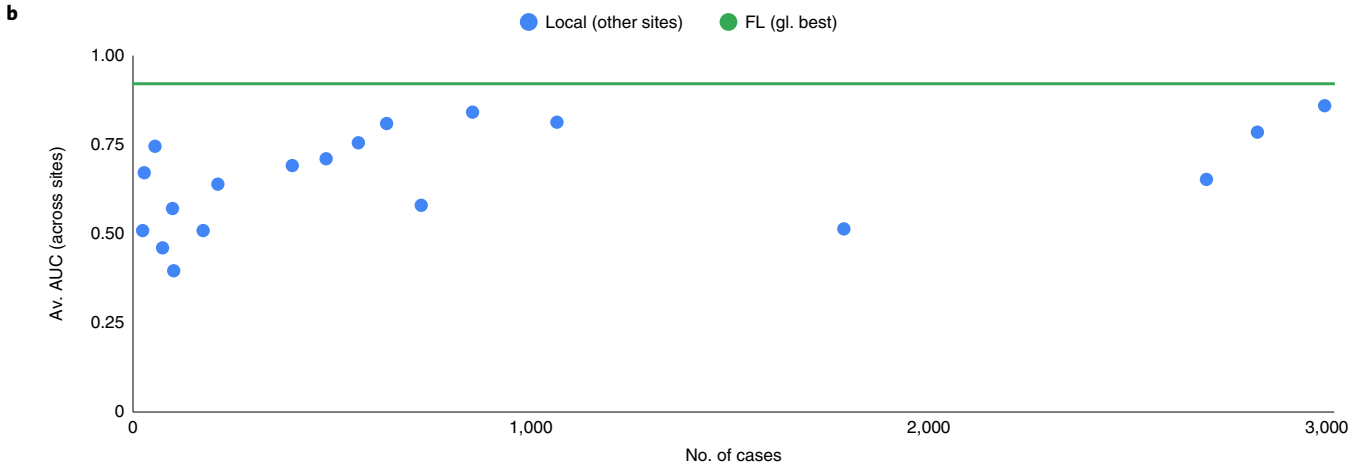


Fig. 2 | Performance of FL versus local models. a, Performance on each client's test set in prediction of 24-h oxygen treatment for models trained on local data only (Local) versus that of the best global model available on the server (FL (gl. best)). Av., average test performance across all sites. **b**, Generalizability (average performance on other sites' test data, as represented by average AUC) as a function of a client's dataset size (no. of cases). The green horizontal line denotes the generalizability performance of the best global model. The performance for 18 of 20 clients is shown, because client 12 had outcomes only for 72-h oxygen (Extended Data Fig. 1) and client 14 had cases only with RA treatment, such that the evaluation metric (av. AUC) was not applicable in either of these cases (Methods). Data for client 14 were also excluded from computation of average generalizability in local models.

clinical informatics circumstances, a meticulous harmonization of the data input was not conducted by the authors (Fig. 1c,d).

We compared locally trained models with the global FL model on each client's test data. Training the model through FL resulted in a significant performance improvement ($P \ll 1 \times 10^{-3}$, Wilcoxon signed-rank test) of 16% (as defined by average AUC when running the model on respective local test sets: from 0.795 to 0.920, or 12.5 percentage points) (Fig. 2a). It also resulted in 38% generalizability improvement (as defined by average AUC when running the model on all test sets: from 0.667 to 0.920, or 25.3 percentage points) of the best global model for prediction of 24-h oxygen treatment compared with models trained only on a site's own data (Fig. 2b). For the prediction results of 72-h oxygen treatment, the best global model training resulted in an average performance improvement of 18% compared to locally trained models, while generalizability of the global model improved on average by 34% (Extended Data Fig. 1). The stability of our results was validated by repeating three runs of local and FL training on different randomized data splits.

Local models that were trained using unbalanced cohorts (for example, mostly mild cases of COVID-19) markedly benefited from the FL approach, with a substantial improvement in prediction average AUC performance for categories with only a few cases. This was evident at client site 16 (an unbalanced dataset), with most patients experiencing mild disease severity and with only a few severe cases. The FL model achieved a higher true-positive rate for the two positive (severe) cases and a markedly lower false-positive rate compared to the local model, both shown in the receiver operating characteristic (ROC) plots and confusion matrices (Fig. 3a and Extended Data Fig. 2). More important, the generalizability of the FL model was considerably increased over the locally trained model.

In the case of client sites with relatively small datasets, the best FL model markedly outperformed not only the local model but also those trained on larger datasets from five client sites in the Boston area of the USA (Fig. 3b).

The global model performed well in predicting oxygen needs at 24/72 h in patients both COVID positive and negative (Extended Data Fig. 3).

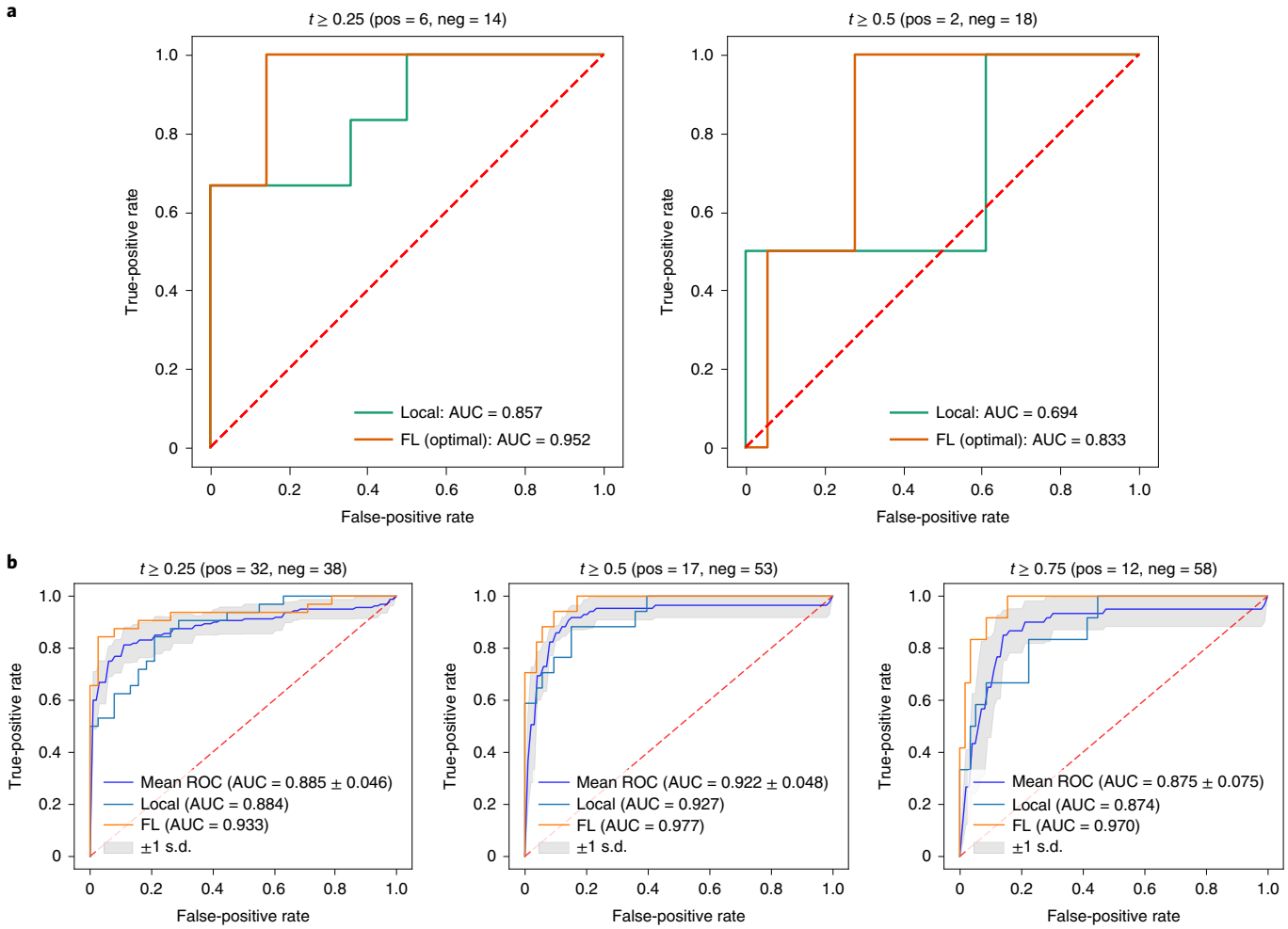


Fig. 3 | Comparison of FL- and locally trained models. a, ROC at client site 16, with unbalanced data and mostly mild cases. **b**, ROC of the local model at client site 12 (a small dataset), mean ROC of models trained on larger datasets corresponding to the five client sites in the Boston area (1, 4, 5, 6, 8) and ROC of the best global model in prediction of 72-h oxygen treatment for different thresholds of EXAM score (left, middle, right). The mean ROC is calculated based on five locally trained models while the gray area denotes the ROC standard deviation. ROCs for three different cutoff values (t) of the EXAM risk score are shown. Pos and neg denote the number of positive and negative cases, respectively, as defined by this range of EXAM score.

Validation at independent sites. Following initial training, EXAM was subsequently tested at three independent validation sites: Cooley Dickinson Hospital (CDH), Martha's Vineyard Hospital (MVH) and Nantucket Cottage Hospital (NCH), all in Massachusetts, USA. The model was not retrained at these sites and it was used only for validation purposes. The cohort size and model inference results are summarized in Table 2, and the ROC curves and confusion matrices for the largest dataset (from CDH) are shown in Fig. 4. The operating point was set to discriminate between nonmechanical ventilation and mechanical ventilation (MV) treatment (or death). The FL global trained model, EXAM, achieved an average AUC of 0.944 and 0.924 for 24- and 72-h prediction tasks, respectively (Table 2), which exceeded the average performance among sites used in training EXAM. For prediction of MV treatment (or death) at 24 h, EXAM achieved a sensitivity of 0.950 and specificity of 0.882 at CDH, and a sensitivity of 1.000 specificity of 0.934 at MVH. NCH did not have any cases with MV/death at 24 h. In regard to 72-h MV prediction, EXAM achieved a sensitivity of 0.929 and specificity of 0.880 at CDH, sensitivity of 1.000 and specificity of 0.976 at MVH and sensitivity of 1.000 and specificity of 0.929 at NCH.

For MV at CDH at 72 h, EXAM had a low false-negative rate of 7.1%. Representative failure cases are presented in Extended Data Fig. 4, showing two false-negative cases from CDH where one case had many missing EMR data features and the other had a CXR with a motion artifact and some missing EMR features.

Use of differential privacy. A primary motivation for healthcare institutes to use FL is to preserve the security and privacy of their data, as well as adherence to data compliance measures. For FL, there remains the potential risk of model 'inversion'[47] or even the reconstruction of training images from the model gradients themselves[48]. To counter these risks, security-enhancing measures were used to mitigate risk in the event of data 'interception' during site-server communication[49]. We experimented with techniques to avoid interception of FL data, and added a security feature that we believe could encourage more institutions to use FL. We thus validated previous findings showing that partial weight sharing, and other differential privacy techniques, can successfully be applied in FL[50]. Through investigation of a partial weight-sharing scheme[50–52], we showed that models can reach a comparable performance even when only 25% of weight updates are shared (Extended Data Fig. 5).

Discussion

This study features a large, real-world healthcare FL study in terms of number of sites and number of data points used. We believe that it provides a powerful proof-of-concept of the feasibility of using

Table 2 | Performance of EXAM on independent datasets. Top, breakdown of patients by level of oxygen required across independent datasets from CDH, MVH and NCH. Bottom, AUC for prediction of the level of oxygen required at 24 and 72 h for the three independent datasets (95% confidence intervals)

Site	Cases (*n*)	Positive cases (*n*)	Prediction interval (h)	Patients at each level of oxygen requirement (*n*)			
				RA	LFO	HFO-NV	MV and death
CDH	840	244	24	608	162	48	22
			72	575	173	62	30
MVH	399	30	24	356	36	3	4
			72	351	39	3	6
NCH	264	29	24	237	23	4	0
			72	235	22	4	3

Site	Prediction interval (h)	≥LFO	≥HFO-NV	≥MV	Average AUC
CDH	24	0.925 (0.903, 0.945)	0.950 (0.926, 0.971)	0.956 (0.918, 0.984)	0.944
	72	0.902 (0.881, 0.924)	0.931 (0.905, 0.955)	0.938 (0.893, 0.927)	0.924
MVH	24	0.904 (0.844, 0.954)	0.836 (0.620, 0.978)	0.964 (0.925, 1.000)	0.901
	72	0.887 (0.827, 0.940)	0.872 (0.663, 0.992)	0.988 (0.973, 0.997)	0.916
NCH	24	0.895 (0.833, 0.950)	0.984 (0.957, 1.000)	N/A	N/A
	72	0.904 (0.850, 0.949)	0.947 (0.890, 0.991)	0.931 (0.897, 0.959)	0.927

The AUC for the NCH dataset for MV at 24 h could not be calculated because there were no mechanically ventilated patients. Cases, number of patients included in the dataset; positive cases, number of patients with confirmed COVID-19 infection included in the dataset; N/A, not available.

FL for fast and collaborative development of needed AI models in healthcare. Our study involved multiple sites across four continents and under the oversight of different regulatory bodies, and thus holds the promise of being provided to different regulated markets in an expedited way. The global FL model, EXAM, proved to be more robust and achieved better results at individual sites than any model trained on only local data. We believe that consistent improvement was achieved owing to a larger, but also a more diverse, dataset, the use of data inputs that can be standardized and avoidance of clinical impressions/reported symptoms. These factors played an important part in increasing the benefits from this FL approach and its impact on performance, generalizability and, ultimately, the model's usability.

For a client site with a relatively small dataset, two typical approaches could be used for fitting a useful model: one is to train locally with its own data, the other is to apply a model trained on a larger dataset. For sites with small datasets, it would have been virtually impossible to build a performant deep learning model using only their local data. The finding, that these two approaches were outperformed on all three prediction tasks by the global FL model, indicates that the benefit for client sites with small datasets arising from participation in FL collaborations is substantial. This is probaby a reflection of FL's ability to capture more diversity than local training, and to mitigate the bias present in models trained on a homogenous population. An under-represented population or age group in one hospital/region might be highly represented in another region—such as children who might be differentially affected by COVID-19, including disease manifestations in lung imaging[46].

The validation results confirmed that the global model is robust, supporting our hypothesis that FL-trained models are generalizable across healthcare systems. They provide a compelling case for the use of predictive algorithms in COVID-19 patient care, and the use of FL in model creation and testing. By participating in this study the client sites received access to EXAM, to be further validated ahead of pursuing any regulatory approval or future introduction into clinical care. Plans are under way to validate EXAM prospectively in 'production' settings at MGB leveraging COVID-19 targeted resources[53], as well as at different sites that were not a part of the EXAM training.

Over 200 prediction models to support decision-making in patients with COVID-19 have been published[19]. Unlike the majority of publications focused on diagnosis of COVID-19 or prediction of mortality, we predicted oxygen requirements that have implications for patient management. We also used cases with unknown SARS-COV-2 status, and so the model could provide input to the physician ahead of receiving a result for PCR with reverse transcription (RT–PCR), making it useful for a real-life clinical setting. The model's imaging input is used in common practice, in contrast with models that use chest computed tomography, a nonconsensual diagnostic modality. The model's design was constrained to objective predictors, unlike many published studies that leveraged subjective clinical impressions. The data collected reflect varied incidence rates, and thus the 'population momentum' we encountered is more diverse. This implies that the algorithm can be useful in populations with different incidence rates.

Patient cohort identification and data harmonization are not novel issues in research and data science[54], but are further complicated, when using FL, given the lack of visibility on other sites' datasets. Improvements to clinical information systems are needed to streamline data preparation, leading to better leverage of a network of sites participating in FL. This, in conjunction with hyperparameter engineering, can allow algorithms to 'learn' more effectively from larger data batches and adapt model parameters to a particular site for further personalization—for example, through further fine-tuning on that site[39]. A system that would allow seamless, close-to real-time model inference and results processing would also be of benefit and would 'close the loop' from training to model deployment.

Because data were not centralized they are not readily accessible. Given that, any future analysis of the results, beyond what was derived and collected, is limited.

Similar to other machine learning models, EXAM is limited by the quality of the training data. Institutions interested in deploying this algorithm for clinical care need to understand potential biases in the training. For example, the labels used as ground truth in the training of the EXAM model were derived from 24- and 72-h oxygen consumption in the patient; it is assumed that oxygen delivered to the patient equates the oxygen need. However, in the early phase of the COVID-19 pandemic, many patients were provided high-flow oxygen prophylactically regardless of their oxygen need. Such clinical practice could skew the predictions made by this model.

Since our data access was limited, we did not have sufficient available information for the generation of detailed statistics regarding failure causes, post hoc, at most sites. However, we did study failure cases from the largest independent test site, CDH, and were able to generate hypotheses that we can test in the future. For high-performing sites, it seems that most failure cases fall into one

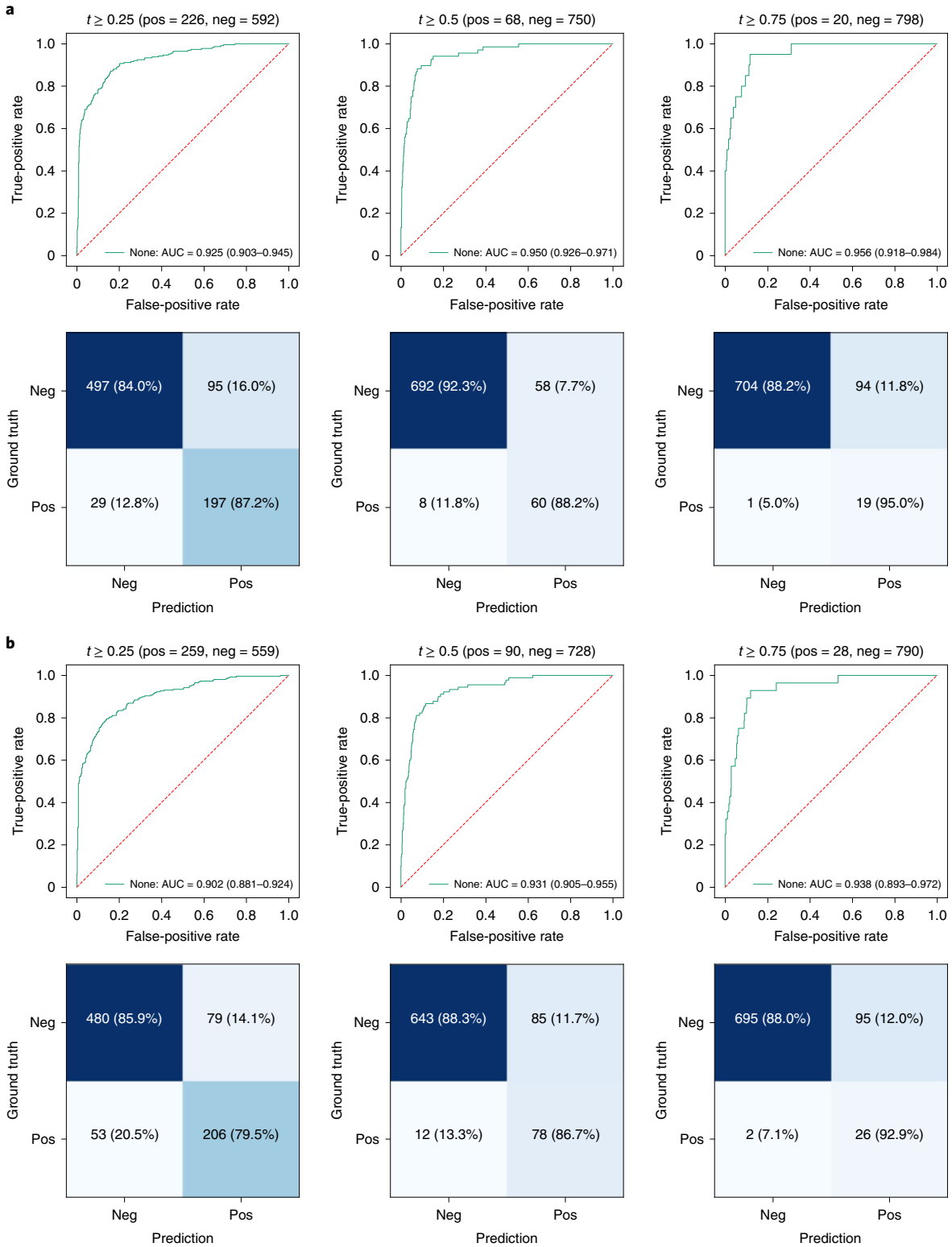


Fig. 4 | Performance of the best global model on the largest independent dataset. a,**b**, Performance (ROC) (top) and confusion matrices (bottom) of the EXAM FL model on the CDH dataset for prediction of oxygen requirement at 24 h (**a**) and 72 h (**b**). ROCs for three different cutoff values (*t*) of the EXAM risk score are shown.

of two categories: (1) low quality of input data—for example, missing data or motion artifact in CXR; or (2) out-of-distribution data—for example a very young patient.

In future, we also intend to investigate the potential for a 'population drift' due to different phases of disease progression. We believe that, owing to the diversity across the 20 sites, this risk may have been mitigated.

A feature that would enhance these kinds of large-scale collaboration is the ability to predict the contribution of each client site towards improving the global FL model. This will help in client

site selection, and in prioritization of data acquisition and annotation efforts. The latter is especially important given the high costs and difficult logistics of these large-consortia endeavors, and it will enable these endeavors to capture diversity rather than the sheer quantity of data samples.

Future approaches may incorporate automated hyperparameter searching[55], neural architecture search[56] and other automated machine learning[57] approaches to find the optimal training parameters for each client site more efficiently.

Known issues of batch normalization (BN) in FL[58] motivated us to fix our base model for image feature extraction[49] to reduce the divergence between unbalanced client sites. Future work might explore different types of normalization techniques to allow the training of AI models in FL more effectively when client data are nonindependent and identically distributed.

Recent works on privacy attacks within the FL setting have raised concerns on data leakage during model training[59]. Meanwhile, protection algorithms remain underexplored and constrained by multiple factors. While differential privacy algorithms[36,48,49] show good protection, they may weaken the model's performance. Encryption algorithms, such as homomorphic encryption[60], maintain performance but may substantially increase message size and training time. A quantifiable way to measure privacy would allow better choices for deciding the minimal privacy parameters necessary while maintaining clinically acceptable performance[36,48,49].

Following further validation, we envision deployment of the EXAM model in the ED setting as a way to evaluate risk at both the per-patient and population level, and to provide clinicians with an additional reference point when making the frequently difficult task of triaging patients. We also envision using the model as a more sensitive population-level metric to help balance resources between regions, hospitals and departments. Our hope is that similar FL efforts can break the data silos and allow for faster development of much-needed AI models in the near future.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-021-01506-3.

Received: 21 December 2020; Accepted: 13 August 2021;
Published online: 15 September 2021

References

1. Budd, J. et al. Digital technologies in the public-health response to COVID-19. *Nat. Med.* **26**, 1183–1192 (2020).
2. Moorthy, V., Henao Restrepo, A. M., Preziosi, M.-P. & Swaminathan, S. Data sharing for novel coronavirus (COVID-19). *Bull. World Health Organ.* **98**, 150 (2020).
3. Chen, Q., Allot, A. & Lu, Z. Keep up with the latest coronavirus research. *Nature* **579**, 193 (2020).
4. Fabbri, F., Bhatia, A., Mayer, A., Schlotter, B. & Kaiser, J. BCG IT spend pulse: how COVID-19 is shifting tech priorities. https://www.bcg.com/publications/2020/how-covid-19-is-shifting-big-it-spend (2020).
5. Candelon, F., Reichert, T., Duranton, S., di Carlo, R. C. & De Bondt, M. The rise of the AI-powered company in the postcrisis world. https://www.bcg.com/en-gb/publications/2020/business-applications-artificial-intelligence-post-covid (2020).
6. Chao, H. et al. Integrative analysis for COVID-19 patient outcome prediction. *Med. Image Anal.* **67**, 101844 (2021).
7. Zhu, X. et al. Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan. *Med. Image Anal.* **67**, 101824 (2021).
8. Yang, D. et al. Federated semi-supervised learning for Covid region segmentation in chest ct using multi-national data from China, Italy, Japan. *Med. Image Anal.* **70**, 101992 (2021).
9. Minaee, S., Kafieh, R., Sonka, M., Yazdani, S. & Jamalipour Soufi, G. Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning. *Med. Image Anal.* **65**, 101794 (2020).
10. COVID-19 Studies from the World Health Organization Database. https://clinicaltrials.gov/ct2/who_table (2020).
11. ACTIV. https://www.nih.gov/research-training/medical-research-initiatives/activ (2020).
12. Coronavirus Treatment Acceleration Program (CTAP). US Food and Drug Administration https://www.fda.gov/drugs/coronavirus-covid-19-drugs/coronavirus-treatment-acceleration-program-ctap (2020).
13. Gleeson, P., Davison, A. P., Silver, R. A. & Ascoli, G. A. A commitment to open source in neuroscience. *Neuron* **96**, 964–965 (2017).
14. Piwowar, H. et al. The state of OA: a large-scale analysis of the prevalence and impact of open access articles. *PeerJ.* **6**, e4375 (2018).
15. European Society of Radiology (ESR). What the radiologist should know about artificial intelligence – an ESR white paper. *Insights Imaging* **10**, 44 (2019).
16. Pesapane, F., Codari, M. & Sardanelli, F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur. Radiol. Exp.* **2**, 35 (2018).
17. Price, W. N. 2nd & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
18. Liang, W. et al. Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern. Med.* **180**, 1081–1089 (2020).
19. Wynants, L. et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *Brit. Med. J.* **369**, m1328 (2020).
20. Zhang, L. et al. D-dimer levels on admission to predict in-hospital mortality in patients with Covid-19. *J. Thromb. Haemost.* **18**, 1324–1329 (2020).
21. Sands, K. E. et al. Patient characteristics and admitting vital signs associated with coronavirus disease 2019 (COVID-19)-related mortality among patients admitted with noncritical illness. https://doi.org/10.1017/ice.2020.461 (2020).
22. American College of Radiology. CR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection. https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection (2020).
23. Rubin, G. D. et al. The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society. *Radiology* **296**, 172–180 (2020).
24. World Health Organization. Use of chest imaging in COVID-19. https://www.who.int/publications/i/item/use-of-chest-imaging-in-covid-19 (2020).
25. Jamil, S. et al. Diagnosis and management of COVID-19 disease. *Am. J. Respir. Crit. Care Med.* **201**, 10 (2020).
26. Redmond, C. E., Nicolaou, S., Berger, F. H., Sheikh, A. M. & Patlas, M. N. Emergency radiology during the COVID-19 pandemic: The Canadian Association of Radiologists Recommendations for Practice. *Can. Assoc. Radiologists J.* **71**, 425–430 (2020).
27. Buch, V. et al. Development and validation of a deep learning model for prediction of severe outcomes in suspected COVID-19 Infection. Preprint at https://arxiv.org/abs/2103.11269 (2021).
28. Lyons, C. & Callaghan, M. The use of high-flow nasal oxygen in COVID-19. *Anaesthesia* **75**, 843–847 (2020).
29. Whittle, J. S., Pavlov, I., Sacchetti, A. D., Atwood, C. & Rosenberg, M. S. Respiratory support for adult patients with COVID-19. *J. Am. Coll. Emerg. Physicians Open* **1**, 95–101 (2020).
30. Ai, J., Li, Y., Zhou, X. & Zhang, W. COVID-19: treating and managing severe cases. *Cell Res.* **30**, 370–371 (2020).
31. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
32. Cahan, E. M., Hernandez-Boussard, T., Thadaney-Israni, S. & Rubin, D. L. Putting the data before the algorithm in big data addressing personalized healthcare. *NPJ Digit. Med.* **2**, 78 (2019).
33. Thrall, J. H. et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *J. Am. Coll. Radiol.* **15**, 504–508 (2018).
34. Shilo, S., Rossman, H. & Segal, E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat. Med.* **26**, 29–38 (2020).
35. Gao, Y. & Cui, Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nat. Commun.* **11**, 5131 (2020).
36. Rieke, N. et al. The future of digital health with federated learning. *NPJ Dig. Med.* **3**, 119 (2020).
37. Yang, Q., Liu, Y., Chen, T. & Tong, Y. Federated machine learning: concept and applications. *ACM Trans. Intell. Syst. Technol.* **10**, 12 (2019).
38. Ma, C. et al. On safeguarding privacy and security in the framework of federated learning. *IEEE Netw.* **34**, 242–248 (2020).
39. Brisimi, T. S. et al. Federated learning of predictive models from federated Electronic Health Records. *Int. J. Med. Inform.* **112**, 59–67 (2018).
40. Roth, H. R. et al. Federated learning for breast density classification: a real-world implementation. In *Proc. Second MICCAI Workshop, DART 2020*

and First MICCAI Workshop, DCL 2020, Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning (eds. Albarqouni, S. et al.) Vol. 12,444, 181–191 (Springer International Publishing, 2020).
41. Sheller, M. J. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 12598 (2020).
42. Remedios, S. W., Butman, J. A., Landman, B. A. & Pham, D. L. in *Federated Gradient Averaging for Multi-Site Training with Momentum-Based Optimizers* (eds Remedios, S. W. et al.) (Springer, 2020).
43. Xu, Y. et al. A collaborative online AI engine for CT-based COVID-19 diagnosis. Preprint at https://www.medrxiv.org/content/10.1101/2020.05.10.20096073v2 (2020).
44. Raisaro, J. L. et al. SCOR: A secure international informatics infrastructure to investigate COVID-19. *J. Am. Med. Inform. Assoc.* **27**, 1721–1726 (2020).
45. Vaid, A. 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**, e24207 (2021).
46. Nino, G. et al. Pediatric lung imaging features of COVID-19: a systematic review and meta-analysis. *Pediatr. Pulmonol.* **56**, 252–263 (2021).
47. Fredrikson, M., Jha, S. & Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proc. 22nd ACM SIGSAC Conference on Computer and Communications Security* 1322–1333, https://doi.org/10.1145/2810103.2813677 (2015).
48. Zhu, L., Liu, Z. & Han, S. in *Advances in Neural Information Processing Systems 32* (eds Wallach, H. et al.) 14774–14784 (Curran Associates, Inc., 2019).
49. Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2**, 305–311 (2020).
50. Li, W. et al. in *Privacy-Preserving Federated Brain Tumour Segmentation* 133–141 (Springer, 2019).
51. Shokri, R. & Shmatikov, V. Privacy-preserving deep learning. In *Proc. 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)* https://doi.org/10.1109/allerton.2015.7447103 (2015).
52. Li, X. et al. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Med. Image Anal.* **65**, 101765 (2020).
53. Estiri, H. et al. Predicting COVID-19 mortality with electronic medical records. *NPJ Dig. Med.* **4**, 15 (2021).
54. Jiang, G. et al. Harmonization of detailed clinical models with clinical study data standards. *Methods Inf. Med.* **54**, 65–74 (2015).
55. Yang, D. et al. in *Searching Learning Strategy with Reinforcement Learning for 3D Medical Image Segmentation*. https://doi.org/10.1007/978-3-030-32245-8_1 (2019).
56. Elsken, T., Metzen, J. H. & Hutter, F. Neural architecture search: a survey. *J. Mach. Learning Res.* **20**, 1–21 (2019).
57. Yao, Q. et al. Taking human out of learning applications: a survey on automated machine learning. Preprint at https://arxiv.org/abs/1810.13306 (2019).
58. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proc. 32nd International Conf. Machine Learning*, PMLR **37**, 448–456 (2015).
59. Kaufman, S., Rosset, S. & Perlich, C. Leakage in data mining: formulation, detection, and avoidance. In *Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 556–563 (2011).
60. Zhang, C. et al. BatchCrypt: efficient homomorphic encryption for cross-silo federated learning. In *Proc. 2020 USENIX Annual Technical Conference, ATC 2020*, 493–506 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

[1]MGH Radiology and Harvard Medical School, Boston, MA, USA. [2]NVIDIA, Santa Clara, CA, USA. [3]Center for Advanced Medical Computing and Analysis, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [4]School of Engineering and Applied Sciences, Harvard University, Boston, MA, USA. [5]San Diego VA Health Care System, San Diego, CA, USA. [6]Department of Neurosurgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [7]Radiology & Imaging Sciences/Clinical Center, National Institutes of Health, Bethesda, MD, USA. [8]National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [9]Division of Cardiovascular Surgery, Department of Surgery, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan. [10]Department of Otolaryngology–Head and Neck Surgery, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan. [11]Graduate Institute of Medical Sciences, National Defense Medical Center, Taipei, Taiwan. [12]Center for Research in Biological Systems, University of California, San Diego, CA, USA. [13]DasaInova, Diagnósticos da América SA, Barueri, Brazil. [14]Division of Pediatric Pulmonary and Sleep Medicine, Children's National Hospital, Washington, DC, USA. [15]Memorial Sloan Kettering Cancer Center, New York, NY, USA. [16]Self-Defense Forces Central Hospital, Tokyo, Japan. [17]Center for Intelligent Imaging, Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, USA. [18]Departments of Radiology and Medical Physics, The University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA. [19]Department of Radiology, NIHR Cambridge Biomedical Resource Centre, University of Cambridge, Cambridge, UK. [20]Department of Internal Medicine, Yeungnam University College of Medicine, Daegu, South Korea. [21]Center for Clinical Data Science, Massachusetts General Brigham, Boston, MA, USA. [22]Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, DC, USA. [23]Departments of Radiology and Pediatrics, The George Washington University School of Medicine and Health Sciences, Washington, DC, USA. [24]Joint Dept. of Medical Imaging, Sinai Health System, University of Toronto, Toronto, Ontario, Canada. [25]Lunenfeld-Tanenbaum Research Institute, Toronto, Ontario, Canada. [26]MeDA Lab Institute of Applied Mathematical Sciences, National Taiwan University, Taipei, Taiwan. [27]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. [28]Research Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand. [29]Center for Artificial Intelligence in Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand. [30]Department of Internal Medicine, School of Medicine, Kyungpook National University, Daegu, South Korea. [31]Departments of Radiology, Medical Physics, and Biomedical Engineering, The University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA. [32]Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand. [33]Thai Red Cross Emerging Infectious Diseases Clinical Center, King Chulalongkorn Memorial Hospital, Bangkok, Thailand. [34]Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA. [35]Department of Radiology, NIHR Cambridge Biomedical Resource Centre, Cambridge University Hospital, Cambridge, UK. [36]Department of Radiology and Imaging Sciences, National Institutes of Health, Bethesda, MD, USA. [37]Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [38]Department of Internal Medicine, Catholic University of Daegu School of Medicine, Daegu, South Korea. [39]Planning and Management Office, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan. [40]School of Medicine, National Defense Medical Center, Taipei, Taiwan. [41]School of Public Health, National Defense Medical Center, Taipei, Taiwan. [42]Graduate Institute of Life Sciences, National Defense Medical Center, Taipei, Taiwan. [43]Medical Review and Pharmaceutical Benefits Division, National Health Insurance Administration, Taipei, Taiwan. [44]Department of Neurosurgery, NYU Grossman School of Medicine, New York, NY, USA. [45]MOST/NTU All Vista Healthcare Center, Center for Artificial Intelligence and Advanced Robotics, National Taiwan University, Taipei, Taiwan. [46]Division of General Internal Medicine and Geriatrics (Fralick), Sinai Health System, Toronto, Ontario, Canada. [47]Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand. [48]Schwartz/Reisman Emergency Medicine Institute, Sinai Health, Toronto, Ontario, Canada. [49]Department of Family and Community Medicine, University of Toronto, Toronto, Ontario, Canada. [50]Department of Medicine and NIHR BioResource for Translational Research, NIHR Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK. [51]Clinical Research Directorate, Frederick National Laboratory for Cancer, National Cancer Institute, Frederick, MD, USA. [52]Department of Microbiology, Sinai Health/University Health Network, Toronto, Ontario, Canada. [53]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [54]Public Health Ontario Laboratories, Toronto, Ontario, Canada. [55]Chulalongkorn University Biomedical Imaging Group and Division of Nuclear Medicine, Department of Radiology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand. [56]These authors contributed equally: Ittai Dayan, Holger Roth, Aoxiao Zhong, Fiona J Gilbert, Quanzheng Li, Mona G. Flores. ✉e-mail: mflores@nvidia.com

Methods

Ethics approval. All procedures were conducted in accordance with the principles for human experimentation as defined in the Declaration of Helsinki and International Conference on Harmonization Good Clinical Practice guidelines, and were approved by the relevant institutional review boards at the following validation sites: CDH, MVH, NCH and at the following training sites: MGB, Mass General Hospital (MGH), Brigham and Women's Hospital, Newton-Wellesley Hospital, North Shore Medical Center and Faulkner Hospital (all eight of these hospitals were covered under MGB's ethics board reference, no. 2020P002673, and informed consent was waived by the instititional review board (IRB). Similarly, participation of the remaining sites was approved by their respective relevant institutional review processes: Children's National Hospital in Washington, DC (no. 00014310, IRB certified exempt); NIHR Cambridge Biomedical Research Centre (no. 20/SW/0140, informed consent waived); The Self-Defense Forces Central Hospital in Tokyo (no. 02-014, informed consent waived); National Taiwan University MeDA Lab and MAHC and Taiwan National Health Insurance Administration (no. 202108026 W, informed consent waived); Tri-Service General Hospital in Taiwan (no. B202105136, informed consent waived); Kyungpook National University Hospital in South Korea (no. KNUH 2020-05-022, informed consent waived); Faculty of Medicine, Chulalongkorn University in Thailand (nos. 490/63, 291/63, informed consent waived); Diagnosticos da America SA in Brazil (no. 26118819.3.0000.5505, informed consent waived); University of California, San Francisco (no. 20-30447, informed consent waived); VA San Diego (no. H200086, IRB certified exempt); University of Toronto (no. 20-0162-C, informed consent waived); National Institutes of Health in Bethesda, Maryland (no. 12-CC-0075, informed consent waived); University of Wisconsin-Madison School of Medicine and Public Health (no. 2016-0418, informed consent waived); Memorial Sloan Kettering Cancer Center in New York (no. 20-194, informed consent waived); and Mount Sinai Health System in New York (no. IRB-20-03271, informed consent waived).

MI-CLAIM guidelines for reporting of clinical AI models were followed (Supplementary Note 2)

Study setting. The study included data from 20 institutions (Fig. 1a): MGB, MGH, Brigham and Women's Hospital, Newton-Wellesley Hospital, North Shore Medical Center and Faulkner Hospital; Children's National Hospital in Washington, DC; NIHR Cambridge Biomedical Research Centre; The Self-Defense Forces Central Hospital in Tokyo; National Taiwan University MeDA Lab and MAHC and Taiwan National Health Insurance Administration; Tri-Service General Hospital in Taiwan; Kyungpook National University Hospital in South Korea; Faculty of Medicine, Chulalongkorn University in Thailand; Diagnosticos da America SA in Brazil; University of California, San Francisco; VA San Diego; University of Toronto; National Institutes of Health in Bethesda, Maryland; University of Wisconsin-Madison School of Medicine and Public Health; Memorial Sloan Kettering Cancer Center in New York; and Mount Sinai Health System in New York. Institutions were recruited between March and May 2020. Dataset curation started in June 2020 and the final data cohort was added in September 2020. Between August and October 2020, 140 independent FL runs were conducted to develop the EXAM model and, by the end of October 2020, EXAM was made public on NVIDIA NGC[61–63]. Data from three independent sites were used for independent validation: CDH, MVH and NCH, all in Massachusetts, USA. These three hospitals had patient population characteristics different from the training sites. The data used for the algorithm validation consisted of patients admitted to the ED at these sites between March 2020 and February 2021, and that satisfied the same inclusion criteria of the data used to train the FL model.

Data collection. The 20 client sites prepared a total of 16,148 cases (both positive and negative) for the purposes of training, validation and testing of the model (Fig. 1b). Medical data were accessed in relation to patients who satisfied the study inclusion criteria. Client sites strived to include all COVID-positive cases from the beginning of the pandemic in December 2019 and up to the time they started local training for the EXAM study. All local training had started by 30 September 2020. The sites also included other patients in the same period with negative RT–PCR test results. Since most of the sites had more SARS-COV-2-negative than -positive patients, we limited the number of negative patients included to, at most, 95% of the total cases at each client site.

A 'case' included a CXR and the requisite data inputs taken from the patient's medical record. A breakdown of the cohort size of the dataset for each client site is shown in Fig. 1b. The distribution and patterns of CXR image intensity (pixel values) varied greatly among sites owing to a multitude of patient- and site-specific factors, such as different device manufacturers and imaging protocols, as shown in Fig. 1c,d. Patient age and EMR feature distribution varied greatly among sites, as expected owing to the differing demographics between globally distributed hospitals (Extended Data Fig. 6).

Patient inclusion criteria. Patient inclusion criteria were: (1) patient presented to the hospital's ED or equivalent; (2) patient had a RT–PCR test performed at any time between presentation to the ED and discharge from the hospital; (3) patient had a CXR in the ED; and (4) patient's record had at least five of the EMR values detailed in Table 1, all obtained in the ED, and the relevant outcomes captured during hospitalization. Of note, The CXR, laboratory results and vitals used were the first available for capture during the visit to the ED. The model did not incorporate any CXR, laboratory results or vitals acquired after leaving the ED.

Model input. In total, 21 EMR features were used as input to the model. The outcome (that is, ground truth) labels were assigned based on patient requirements after 24- and 72-h periods from initial admission to the ED. A detailed list of the requested EMR features and outcomes can be seen in Table 1.

The distribution of oxygen treatment using different devices at different client sites is shown in Extended Data Fig. 7, which details the device usage at admission to the ED and after 24- and 72-h periods. The difference in dataset distribution between the largest and smallest client sites can be seen in Extended Data Fig. 8.

The number of positive COVID-19 cases, as confirmed by a single RT–PCR test obtained at any time between presentation to the ED and discharge from the hospital, is listed in Supplementary Table 1. Each client site was asked to randomly split its dataset into three parts: 70% for training, 10% for validation and 20% for testing. For both 24- and 72-h outcome prediction models, random splits for each of the three repeated local and FL training and evaluation experiments were independently generated.

EXAM model development. There is wide variation in the clinical course of patients who present to hospital with symptoms of COVID-19, with some experiencing rapid deterioration in respiratory function requiring different interventions to prevent or mitigate hypoxemia[62,63]. A critical decision made during the evaluation of a patient at the initial point of care, or in the ED, is whether the patient is likely to require more invasive or resource-limited countermeasures or interventions (such as MV or monoclonal antibodies), and should therefore receive a scarce but effective therapy, a therapy with a narrow risk–benefit ratio due to side effects or a higher level of care, such as admittance to the intensive care unit[64]. In contrast, a patient who is at lower risk of requiring invasive oxygen therapy may be placed in a less intensive care setting such as a regular ward, or even released from the ED for continuing self-monitoring at home[65]. EXAM was developed to help triage such patients.

Of note, the model is not approved by any regulatory agency at this time and it should be used only for research purposes.

EXAM score. EXAM was trained using FL; it outputs a risk score (termed EXAM score) similar to CORISK[27] (Extended Data Fig. 9a) and can be used in the same way to triage patients. It corresponds to a patient's oxygen support requirements within two windows—24 and 72 h—after initial presentation to the ED. Extended Data Fig. 9b illustrates how CORISK and the EXAM score can be used for patient triage.

Chest X-ray images were preprocessed to select the anterior position image and exclude lateral view images, and then scaled to a resolution of 224 × 224. As shown in Extended Data Fig. 9a, the model fuses information from both EMR and CXR features (based on a modified ResNet34 with spatial attention[66] pretrained on the CheXpert dataset)[67] and the Deep & Cross network[68]. To converge these different data types, a 512-dimensional feature vector was extracted from each CXR image using a pretrained ResNet34, with spatial attention, then concatenated with the EMR features as the input for the Deep & Cross network. The final output was a continuous value in the range 0–1 for both 24- and 72-h predictions, corresponding to the labels described above, as shown in Extended Data Fig. 9b. We used cross-entropy as the loss function and 'Adam' as the optimizer. The model was implemented in Tensorflow[69] using the NVIDIA Clara Train SDK[70]. The average AUC for the classification tasks (≥LFO, ≥HFO/NIV or ≥MV) was calculated and used as the final evaluation metric, with normalization to zero-mean and unit variance. CXR images were preprocessed to select the correct series and exclude lateral view images, then scaled to a resolution of 224 × 224 (ref. [27]).

Feature imputation and normalization. A MissForest algorithm[71] was used to impute EMR features, based on the local training dataset. If an EMR feature was completely missing from a client site dataset, the mean value of that feature, calculated exclusively on data from MGB client sites, was used. Then, EMR features were rescaled to zero-mean and unit variance based on statistics calculated on data from the MGB client sites.

Details of EMR–CXR data fusion using the Deep & Cross network. To model the interactions of features from EMR and CXR data at the case level, a deep-feature scheme was used based on a Deep & Cross network architecture[68]. Binary and categorical features for the EMR inputs, as well as 512-dimensional image features in the CXR, were transformed into fused dense vectors of real values by embedding and stacking layers. The transformed dense vectors served as input to the fusion framework, which specifically employed a crossing network to enforce fusion among input from different sources. The crossing network performed explicit feature crossing within its layers, by conducting inner products between the original input feature and output from the previous layer, thus increasing the degree of interaction across features. At the same time, two individual classic deep neural networks with several stacked, fully connected feed-forward layers

were trained. The final output of our framework was then derived from the concatenation of both classic and crossing networks.

FL details. Arguably the most established form of FL is implemention of the federated averaging algorithm as proposed by McMahan et al.[72], or variations thereof. This algorithm can be realized using a client-server setup where each participating site acts as a client. One can think of FL as a method aiming to minimize a global loss function by reducing a set of local loss functions, which are estimated at each site. By minimizing each client site's local loss while also synchronizing the learned client site weights on a centralized aggregation server, one can minimize global loss without needing to access the entire dataset in a centralized location. Each client site learns locally, and shares model weight updates with a central server that aggregates contributions using secure sockets layer encryption and communication protocols. The server then sends an updated set of weights to each client site after aggregation, and sites resume training locally. The server and client site iterate back and forth until the model converges (Extended Data Fig. 9c).

A pseudoalgorithm of FL is shown in Supplementary Note 1. In our experiments, we set the number of federated rounds at $T=200$, with one local training epoch per round t at each client. The number of clients, K, was up to 20 depending on the network connectivity of clients or available data for a specific targeted outcome period (24 or 72 h). The number of local training iterations, n_k, depends on the dataset size at each client k and is used to weigh each client's contributions when aggregating the model weights in federated averaging. During the FL training task, each client site selects its best local model by tracking the model's performance on its local validation set. At the same time, the server determines the best global model based on the average validation scores sent from each client site to the server after each FL round. After FL training finishes, the best local models and the best global model are automatically shared with all client sites and evaluated on their local test data.

When training on local data only (the baseline), we set the epoch number to 200. The Adam optimizer was used for both local training and FL with an initial learning rate of 5×10^{-5} and a stepwise learning rate decay with a factor 0.5 after every 40 epochs, which is important for the convergence of federated averaging[73]. Random affine transformations, including rotation, translations, shear, scaling and random intensity noise and shifts, were applied to the images for data augmentation during training.

Owing to the sensitivity of BN layers[58] when dealing with different clients in a nonindependent and identically distributed setting, we found the best model performance occurred when keeping the pretrained ResNet34 with spatial attention[47] parameters fixed during FL training (that is, using a learning rate of zero for those layers). The Deep & Cross network that combines image features with EMR features does not contain BN layers and hence was not affected by BN instability issues.

In this study we investigated a privacy-preserving scheme that shares only partial model updates between server and client sites. The weight updates were ranked during each iteration by magnitude of contribution, and only a certain percentage of the largest weight updates was shared with the server. To be exact, weight updates (also known as gradients) were shared only if their absolute value was above a certain percentile threshold, k(t) (Extended Data Fig. 5), which was computed from all non-zero gradients, $\Delta W_k^{(t)}$, and could be different for each client k in each FL round t. Variations of this scheme could include additional clipping of large gradients or differential privacy schemes[49] that add random noise to the gradients, or even to the raw data, before feeding into the network[51].

Statistical analysis. We conducted a Wilcoxon signed-rank test to confirm the significance of the observed improvement in performance between the locally trained model and the FL model for the 24- and 72-h time points (Fig. 2 and Extended Data Fig. 1). The null hypothesis was rejected with one-sided $P \ll 1\times10^{-3}$ in both cases.

Pearson's correlation was used to assess the generalizability (robustness of the average AUC value to other client sites' test data) of locally trained models in relation to respective local dataset size. Only a moderate correlation was observed ($r=0.43$, $P=0.035$, degrees of freedom (df) = 17 for the 24-h model and $r=0.62$, $P=0.003$, df = 16 for the 72-h model). This indicates that dataset size alone is not the only factor determining a model's robustness to unseen data.

To compare ROC curves from the global FL model and local models trained at different sites (Extended Data Fig. 3), we bootstrapped 1,000 samples from the data and computed the resulting AUCs. We then calculated the difference between the two series and standardized using the formula $D=(\mathrm{AUC1}-\mathrm{AUC2})/s$, where D is the standardized difference, s is the standard deviation of the bootstrap differences and AUC1 and AUC2 are the corresponding bootstrapped AUC series. By comparing D with normal distribution, we obtained the P values illustrated in Supplementary Table 2. The results show that the null hypothesis was rejected with very low P values, indicating the statistical significance of the superiority of FL outcomes. The computation of P values was conducted in R with the pROC library[74].

Since the model predicts a discrete outcome, a continuous score from 0 to 1, a straightforward calibration evaluation such as a qqplot is not possible. Hence, for a quantified estimate of calibration we quantified discrimination (Extended Data Fig. 10). We conducted one-way analysis of variation (ANOVA) tests to compare local and FL model scores among four ground truth categories (RA, LFO, HFO, MV). The F-statistic, calculated as the variation between the sample means divided by variation within the samples and representing the degree of dispersion among different groups, was used to quantify the models. Our results show that the F-values of five different local sites are 245.7, 253.4, 342.3, 389.8 and 634.8, while that of the FL model is 843.5. Given that larger F-values mean that groups are more separable, the scores from our FL model clearly show a greater dispersion among the four ground truth categories. Furthermore, the P value of the ANOVA test on the FL model is $<2\times10^{-16}$, indicating that the FL prediction scores are statistically significantly different among the different prediction classes.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The dataset from the 20 institutes that participated in this study remains under their custody. These data were used for training at each of the local sites and were not shared with any of the other participating institutions or with the federated server, and they are not publicly available. Data from the independent validation sites are maintained by CAMCA, and access can be requested by contacting Q.L. Based on determination by CAMCA, a data-sharing review and amendment of IRB for research purposes can be conducted by MGB research administration and in accordance with MGB IRB and policy.

Code availability

All code and software used in this study are publicly available at NGC. To access, log in as a guest or create a profile then enter one of the URLs below. The trained models, data preparation guidelines, code for training, validating testing of the model, readme file, installation guideline and license files are publicly available at NVIDIA NGC[61]: https://ngc.nvidia.com/catalog/models/nvidia:med:clara_train_covid19_exam_ehr_xray The federated learning software is available as part of the Clara Train SDK: https://ngc.nvidia.com/catalog/containers/nvidia:clara-train-sdk. Alternatively, use this command to download the model "wget --content-disposition https://api.ngc.nvidia.com/v2/models/nvidia/med/clara_train_covid19_exam_ehr_xray/versions/1/zip -O clara_train_covid19_exam_ehr_xray_1.zip".

References

61. *Nvidia NGC Catalog: COVID-19 Related Models*. https://ngc.nvidia.com/catalog/models?orderBy=scoreDESC&pageNumber=0&query=covid&quickFilter=models&filters (2020).
62. Marini, J. J. & Gattinoni, L. Management of COVID-19 respiratory distress. *JAMA* **323**, 2329–2330 (2020).
63. Cook, T. M. et al. Consensus guidelines for managing the airway in patients with COVID-19: Guidelines from the Difficult Airway Society, the Association of Anaesthetists the Intensive Care Society, the Faculty of Intensive Care Medicine and the Royal College of Anaesthetist. *Anaesthesia* **75**, 785–799 (2020).
64. Galloway, J. B. et al. A clinical risk score to identify patients with COVID-19 at high risk of critical care admission or death: an observational cohort study. *J. Infect.* **81**, 282–288 (2020).
65. Kilaru, A. S. et al. Return hospital admissions among 1419 COVID-19 patients discharged from five U.S. emergency departments. *Acad. Emerg. Med.* **27**, 1039–1042 (2020).
66. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* https://doi.org/10.1109/cvpr.2016.90 (2016).
67. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. AAAI Conf. Artif. Intell.* **33**, 590–597 (2019).
68. Wang, R., Fu, B., Fu, G. & Wang, M. Deep & Cross network for Ad Click predictions. In *Proc. ADKDD'17* Article no. 12 (2017).
69. Abadi, M. et al. TensorFlow: asystem for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, USENIX Association 265–283 (2016).
70. *NVIDIA Clara Imaging*. https://developer.nvidia.com/clara-medical-imaging (2020).
71. Stekhoven, D. J. & Bühlmann, P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
72. McMahan, H., Moore, E., Ramage, D., Hampson, S. & y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. http://proceedings.mlr.press/v54/mcmahan17a.html (2017).
73. Hsieh, K., Phanishayee, A., Mutlu, O. & Gibbons, P. B. The non-IID data quagmire of decentralized machine learning. In *Proc. 37th International Conf. Machine Learning* PMLR 119 (2020).
74. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

Acknowledgements

The views expressed in this study are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care or any of the organizations associated with the authors. MGB thank the following individuals for their support: J. Brink, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA; M. Kalra, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA; N. Neumark, Center for Clinical Data Science, Massachusetts General Brigham, Boston, MA; T. Schultz, Department of Radiology, Massachusetts General Hospital, Boston, MA; N. Guo, Center for Advanced Medical Computing and Analysis, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA; J. K. Cramer, Director, QTIM lab at the Athinoula A. Martinos Center for Biomedical Imaging at MGH; S. Pomerantz, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA; G. Boland, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; W. Mayo-Smith, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA. UCSF thank P. B. Storey, J. Chan and J. Block for implementing the UCSF FL client infrastructure, and W. Tellis for providing the source imaging repository for this work. The UCSF EMR and clinical notes for this study were accessed via the COVID-19 Research Data Mart, https://data.ucsf.edu/covid19. The Faculty of Medicine, Chulalongkorn University thank the Ratchadapisek Sompoch Endowment Fund RA (PO) (no. 001/63) for the collection and management of COVID-19-related clinical data and biological specimens for the Research Task Force, Faculty of Medicine, Chulalongkorn University. NIHR Cambridge Biomedical Research Centre thank A. Priest, who is supported by the NIHR (Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust). National Taiwan University MeDA Lab and the MAHC and Taiwan National Health Insurance Administration thank the MOST Joint Research Center for AI technology, the All Vista Healthcare National Health Insurance Administration, Taiwan, the Ministry of Science and Technology, Taiwan and the National Center for Theoretical Sciences Mathematics Division. National Institutes of Health (NIH) acknowledge that the NIH Medical Research Scholars Program is a public–private partnership supported jointly by the NIH and by generous contributions to the Foundation for the NIH from the Doris Duke Charitable Foundation, the American Association for Dental Research, the Colgate-Palmolive Company, Genentech, alumni of student research programs and other individual supporters via contributions to the Foundation for the NIH.

Author contributions

I.D. and M.G.F. contributed to the acquisition of data, study support, drafting and revising the manuscript, study design, study concept and analysis and interpretation of the data. H.R.R., A.Z. and Q.L. contributed to the acquisition of data, study support, drafting and revising the manuscript, study design and analysis and interpretation of data. F.J.G. contributed to the acquisition of data, study support and drafting and revising the manuscript. J.G. contributed to support of the study, drafting and revising the manuscript and analysis and interpretation of data. V.B. contributed to the acquisition of data, study support and study design. D.X. contributed to the acquisition of data, study support, drafting and revising the manuscript and analysis and interpretation of data. A.B.C., B.J.W., J.W.G. and K.J. contributed to the acquisition of data and drafting, and revising the manuscript. N.R. contributed to the support of the study and drafting and revising the manuscript. A.H., A.Z.A., A.L., C.K.L., P.R., E.H., G.L., J.T., K.K., P.M.C.e.S., A.Q., A.F., C.C., D.B., I.Y., M.A. and Y.W. contributed to the support of the study. A.G., C.-S.T., C.-H.W., C.-N.H., D.W., F.C.K., G.C.d.A.C., G.N., H.-H.S., H.O., H.R.R., J.C.C., J.D.K., J.G.P., K.D., M.A.B.C.R., M.G.L., M.A.H., M.A., P.F.D., P.W., S.X., S.K., S.S., S.Y.P., T.M.G., W.J., W.W., W.Y.T., X.L., X.L., Y.J.K., A.N.P., B.T., B.G., B.B., B.S.K., C.T.-D., C.-C.L., C.-J.H., C.L., C.-L.L., C.P.H., E.K.O., E.L., H.S., H.M., J.H.S., K.N.K.M., L.-C.F., M.R.F.d,M., M.F., M.K.K., N.G., P.V., P.E., S.H., S.M., S.L.M., S.R., S.G., S.H., T.K., T.P., T.M., V.L.d.L., Y.R. and Y.R.L. contributed to the acquisition of data.

Competing interests

This study was organized and coordinated by NVIDIA. Y.W., M.A., I.Y., A.Q., C.C., D.B., A.F., H.R.R., J.G., D.X., N.R., A.H., K.K., C.R., A.A., C.K.L., E.H., A.L., G.L., P.M.C.S., J.T. and M.G.F. are employees of NVIDIA and own stock as part of the standard compensation package. J.G. declares ownership of NVIDIA stock. I.D. is presently an officer and shareholder of Rhino HealthTech, Inc., which provides systems for distributed computation that can, among other things, be used to complete FL tasks. He was not employed by this company during the execution of the EXAM study. The remaining authors declare no competing interests. C.H. declares research travel with Siemens Healthineers AG; conference travel; and EUROKONGRESS and personal fees (consultant to GE Healthcare LLC, DSMB member, Focused Ultrasound Foundation). F.J.G. declares research collaboration with Merantix, Screen-Point, Lunit, Volpara and GE Healthcare, and undertakes paid consultancy for Kheiron and Alphabet. M.L. declares that he is the cofounder of PediaMetrix, Inc. and is on the board of the SIPAIM Foundation. S.E.H. declares research collaborations with Merantix, Screen-Point, Lunit and Volpara. B.J.W. and S.X. declare that NIH and NVIDIA have a Cooperative Research and Development Agreement. This work was supported (in part) by the NIH Center for Interventional Oncology and the Intramural Research Program of the National Institutes of Health, via intramural NIH grant nos. Z1ACL040015 and 1ZIDBC011242. Work was supported by the NIH Intramural Targeted Anti-COVID-19 Program, funded by the National Institute of Allergy and Infectious Diseases. NIH may have intellectual property in the field. M.F. is a consultant for Proof Diagnostics, a start-up company developing a CRISPR-based diagnostic test for COVID-19. The remaining authors declare no competing interests.

Additional information

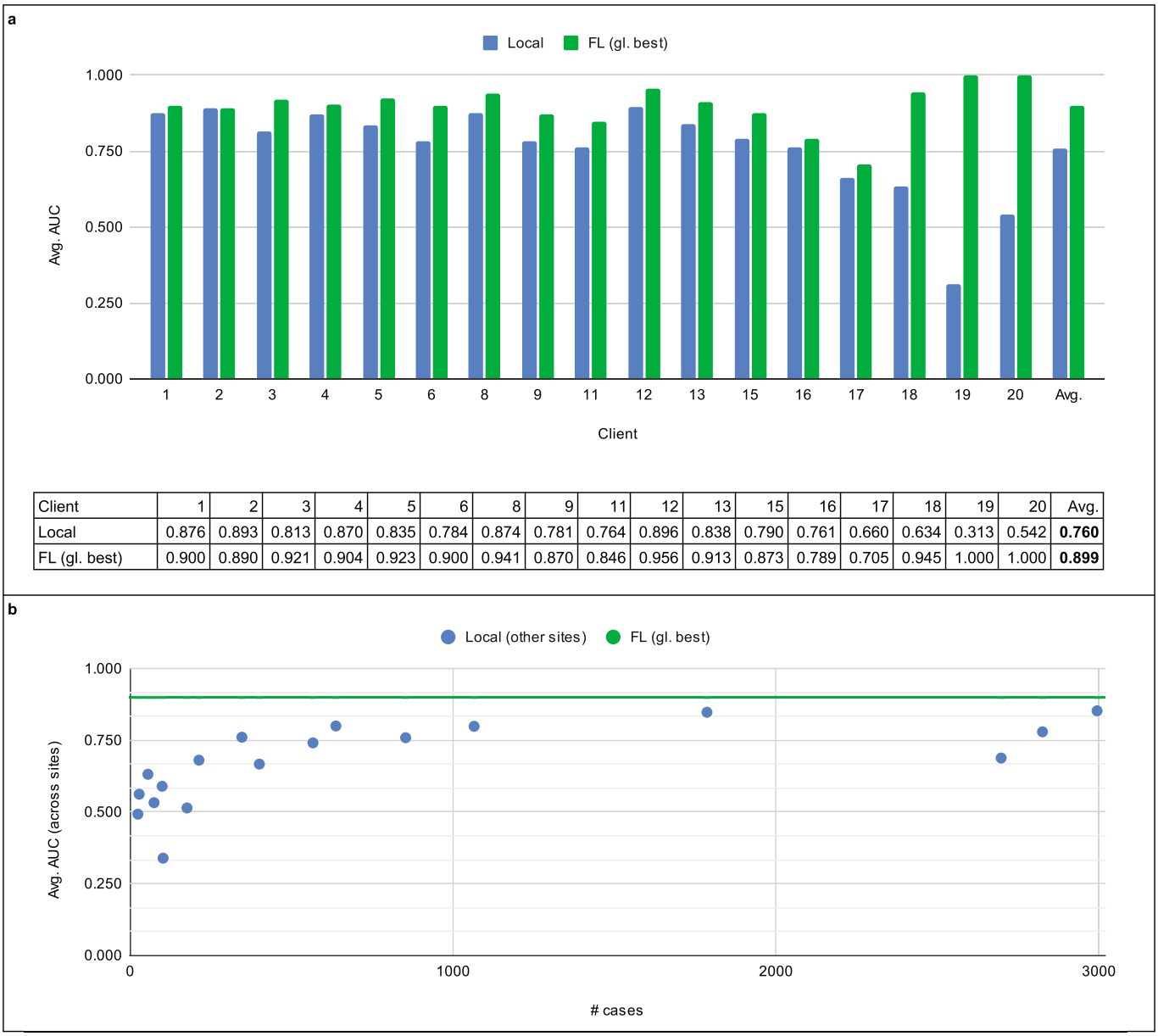
Extended data is available for this paper at https://doi.org/10.1038/s41591-021-01506-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-021-01506-3.

Correspondence and requests for materials should be addressed to Mona G. Flores.

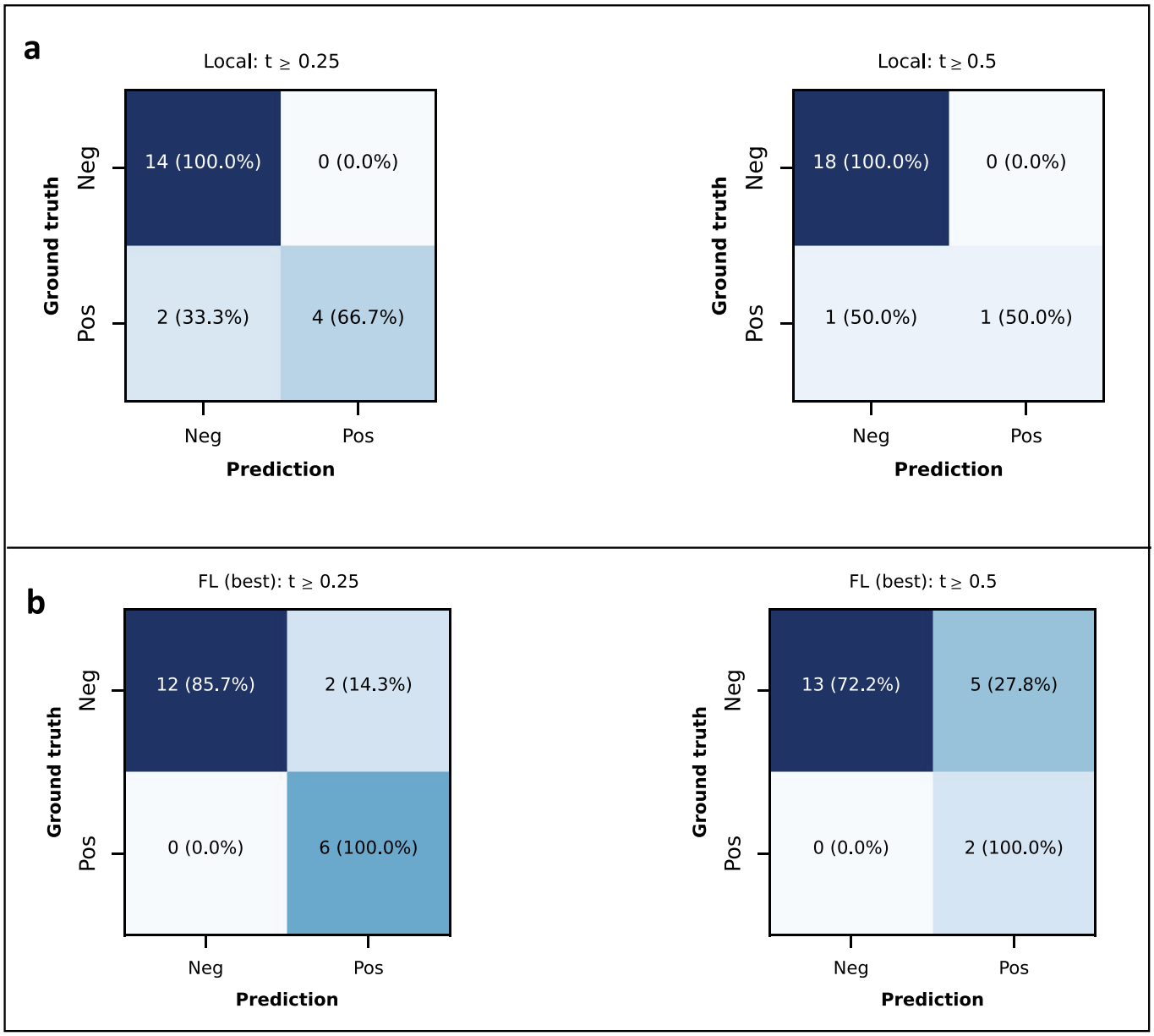
Peer review information *Nature Medicine* thanks Fei Wang, Nikos Paragios and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Michael Basson was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

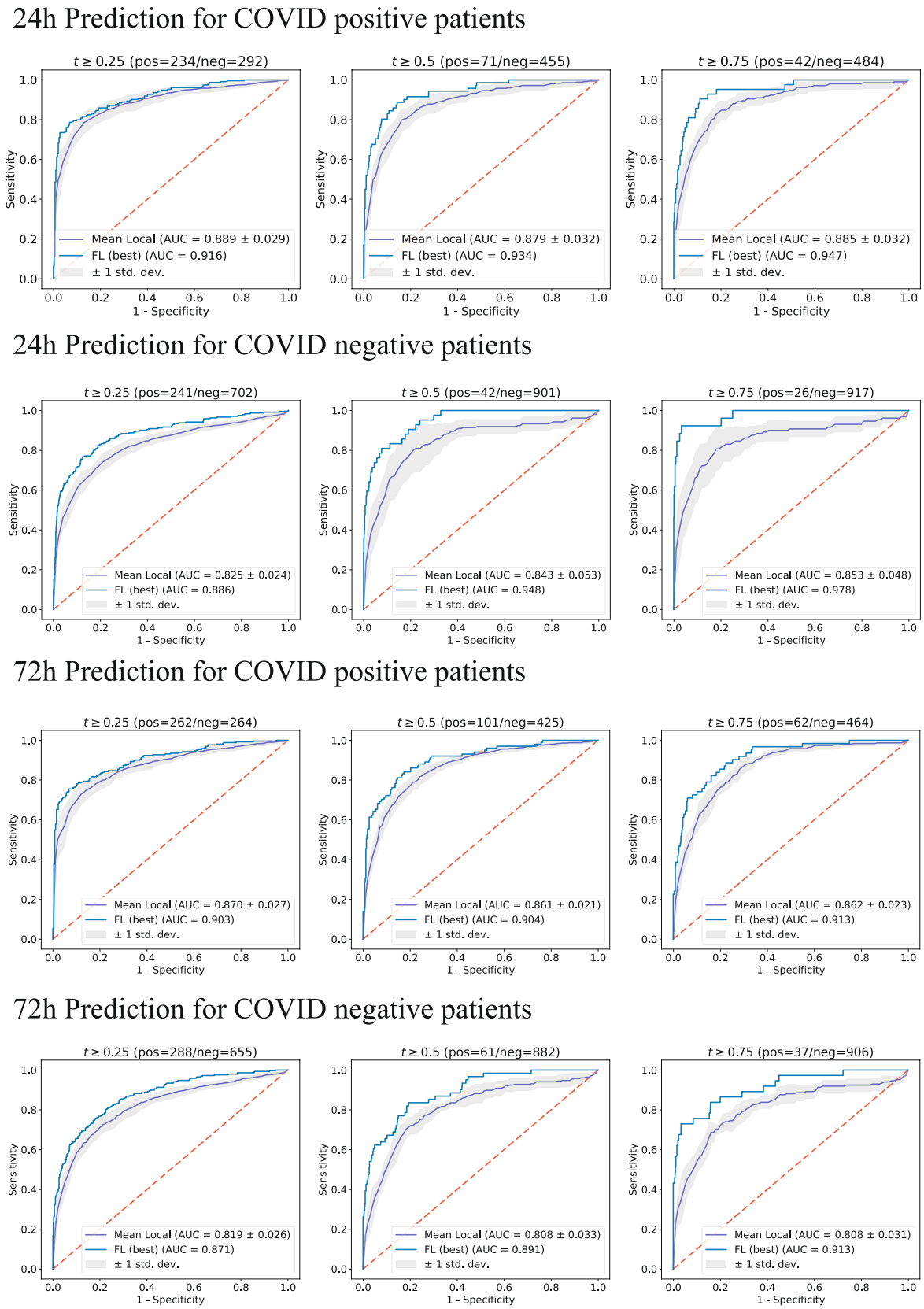


Client	1	2	3	4	5	6	8	9	11	12	13	15	16	17	18	19	20	Avg.
Local	0.876	0.893	0.813	0.870	0.835	0.784	0.874	0.781	0.764	0.896	0.838	0.790	0.761	0.660	0.634	0.313	0.542	**0.760**
FL (gl. best)	0.900	0.890	0.921	0.904	0.923	0.900	0.941	0.870	0.846	0.956	0.913	0.873	0.789	0.705	0.945	1.000	1.000	**0.899**

Extended Data Fig. 1 | Test performance of models predicting 72 h oxygen treatment trained on local data only versus the performance of the best global model available on the server. Test performance of models predicting 72 h oxygen treatment trained on local data only (Local) versus the performance of the best global model available on the server (FL (gl. best)). b, Generalizability (average performance on other sites' test data) as a function of a site's dataset size (# cases). The average performance improved by 18% (from 0.760 to 0.899 or 13.9 percentage points) compared to locally trained models alone, while average generalizability of the global model improved by 34% (from 0.669 to 0.899 or 23.0 percentage points).



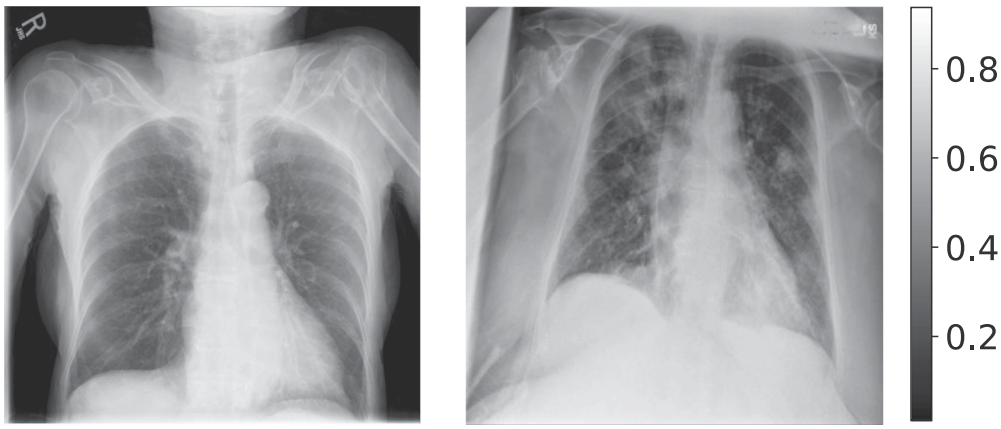
Extended Data Fig. 2 | Confusion Matrices at a site with unbalanced data and mostly mild cases. Confusion Matrices at a site with unbalanced data and mostly mild cases. **a**, Confusion matrices on the test data at site 16 predicting oxygen treatment at 72 h using the locally trained model. **b**, Confusion matrices on the test data at site 16 predicting oxygen treatment at 72 h using the best Federated Learning global model. We show the ROCs for two different cut-off values t of the EXAM risk score.



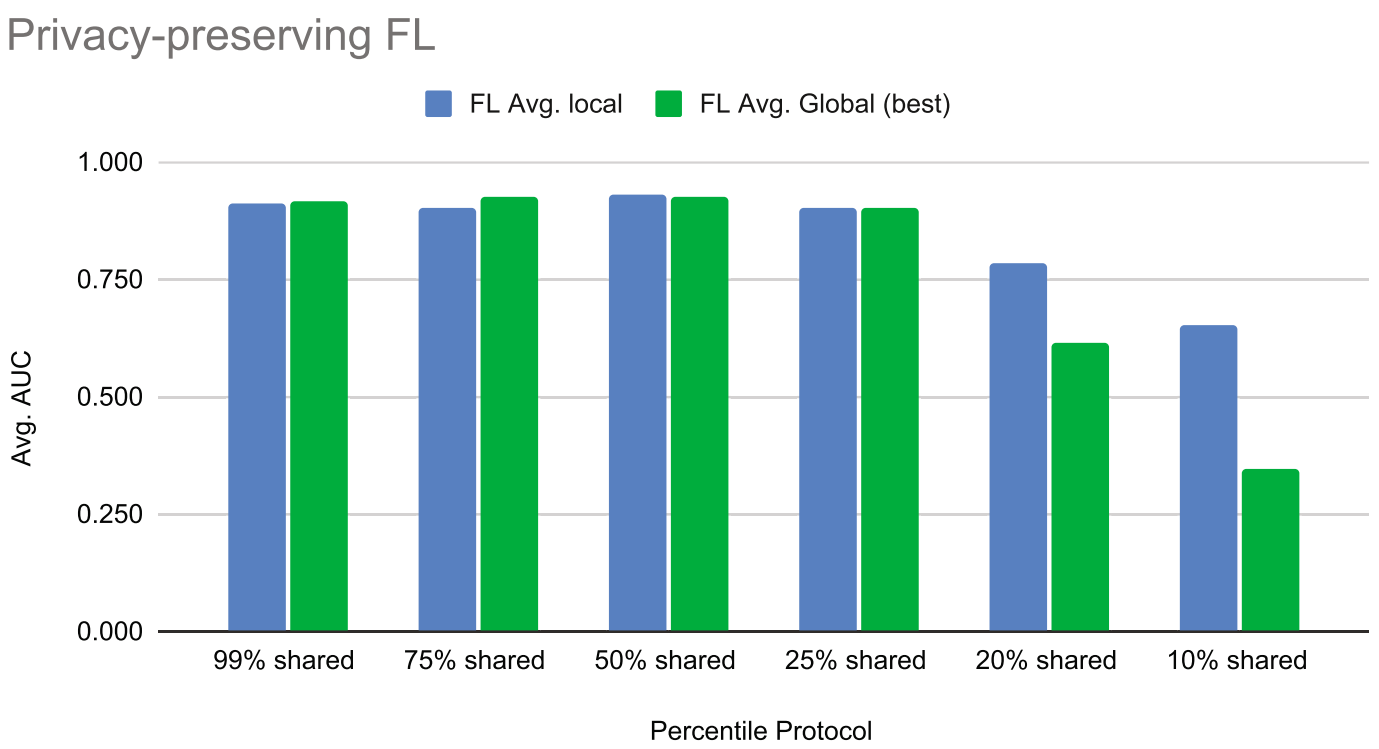
Extended Data Fig. 3 | Model performance on COVID-positive and COVID-negative patients. ROCs of the best global model in comparison to the mean ROCs of models trained on local datasets to predict 24/72-h oxygen treatment devices for COVID positive/negative patients respectively, using the test data of 5 large datasets from sites in the Boston area. The Mean ROC is calculated based on 5 locally trained models, with the gray-area showing the standard deviation of the ROCs. We show the ROCs for three different cut-off values t of the EXAM risk score.

FEAT_VITAL_DBP_FIRST: 54.0
FEAT_VITAL_SBP_FIRST: 136.0
FEAT_PT_AGE: 87
FEAT_LAB_LDH_FIRST: NaN
FEAT_LAB_CRP_FIRST: NaN
FEAT_VITAL_SPO2_FIRST: 97.0
FEAT_VITAL_RR_FIRST: 17.0
FEAT_LAB_AST_FIRST: 26.0
FEAT_LAB_PCLC_FIRST: NaN
FEAT_LAB_LAC_FIRST: NaN
FEAT_LAB_NEUT_FIRST: 4.23
FEAT_LAB_GLU_FIRST: 79.0
FEAT_LAB_WBC_FIRST: 6.34
FEAT_LAB_TNT_FIRST: 16.0
FEAT_LAB_GFR_FIRST: 45.0
FEAT_LAB_CR_FIRST: 1.1
FEAT_LAB_DDMR_FIRST: NaN
FEAT_ED_OD: RA
PCR POS ED: True
PCR POS EVER : True

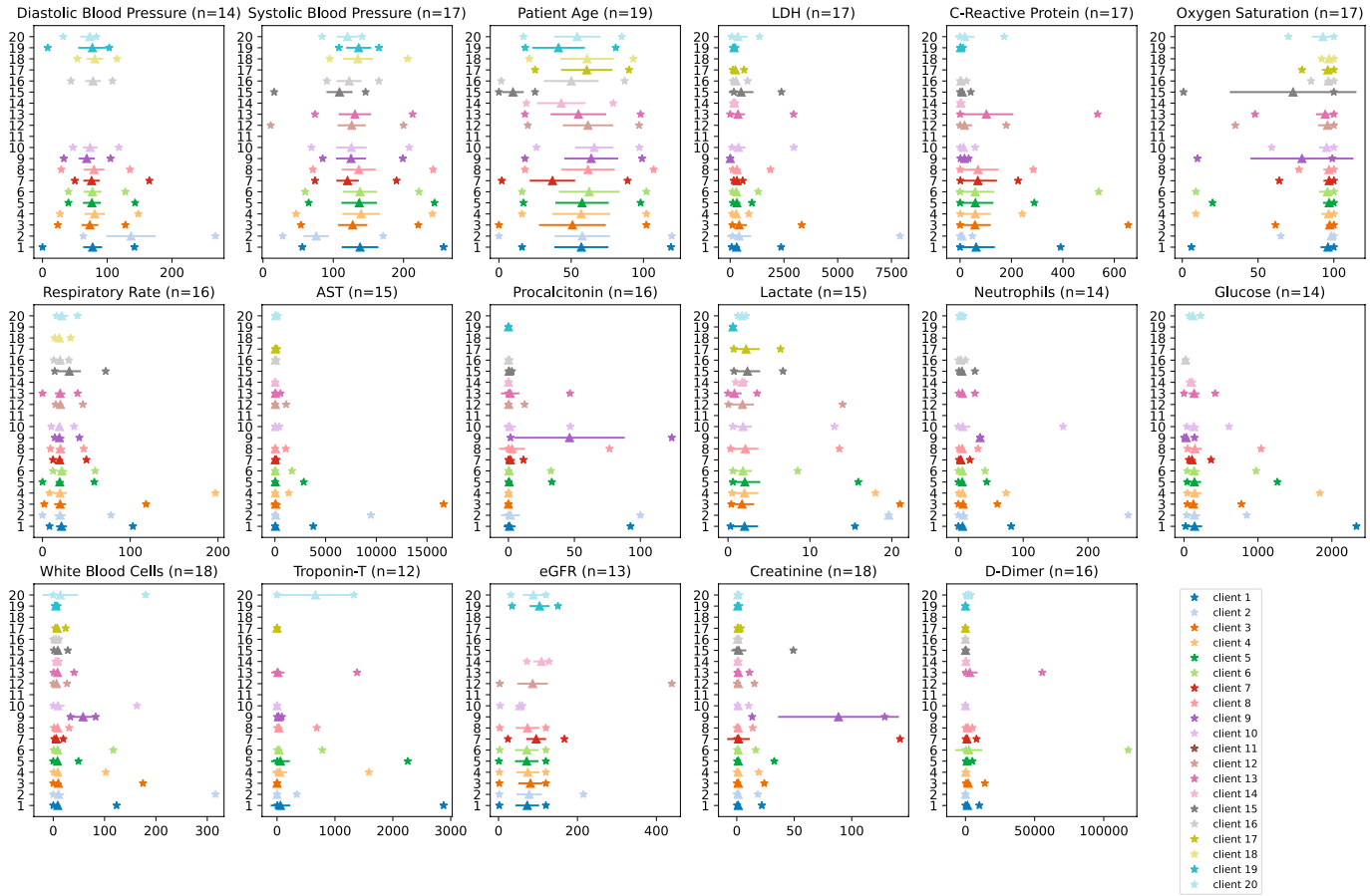
FEAT_VITAL_DBP_FIRST: 84.0
FEAT_VITAL_SBP_FIRST: 146.0
FEAT_PT_AGE: 72
FEAT_LAB_LDH_FIRST: 228.0
FEAT_LAB_CRP_FIRST: 102.0
FEAT_VITAL_SPO2_FIRST: 94.0
FEAT_VITAL_RR_FIRST: 16.0
FEAT_LAB_AST_FIRST: 4.0
FEAT_LAB_PCLC_FIRST: NaN
FEAT_LAB_LAC_FIRST: 1.09
FEAT_LAB_NEUT_FIRST: 6.63
FEAT_LAB_GLU_FIRST: 165.0
FEAT_LAB_WBC_FIRST: 10.52
FEAT_LAB_TNT_FIRST: NaN
FEAT_LAB_GFR_FIRST: 47.0
FEAT_LAB_CR_FIRST: 1.2
FEAT_LAB_DDMR_FIRST: NaN
FEAT_ED_OD: RA
PCR POS ED: True
PCR POS EVER : True



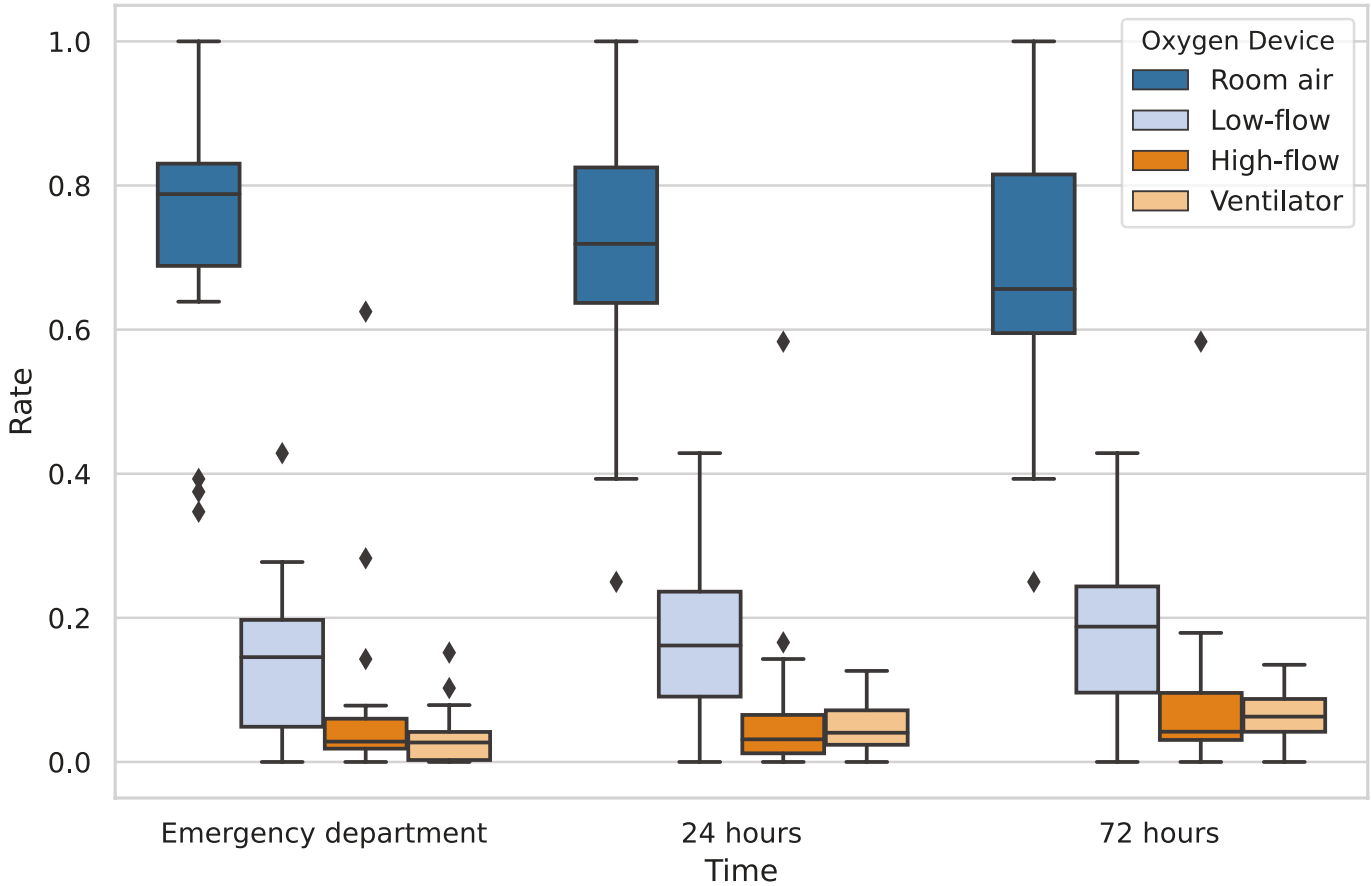
Extended Data Fig. 4 | Failures cases at an independent test site. Failures cases at an independent test site. CXRs from two failure cases at CDH. The above is noisy data where each available value has been anonymized by adding a zero-mean Gaussian noise with the standard deviation of 1/5 of the standard deviation of the cohort distribution.



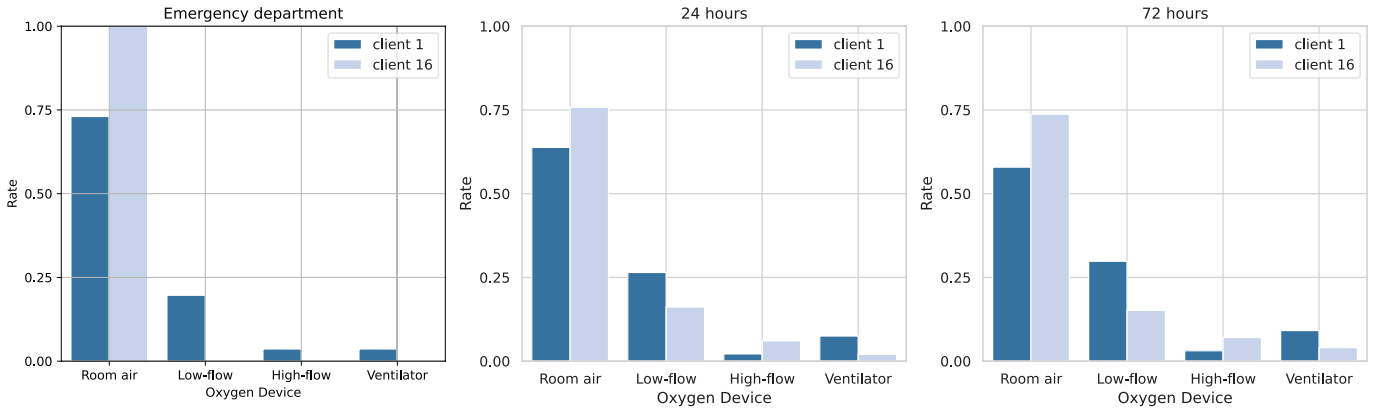
Extended Data Fig. 5 | Safety enhancing features used in EXAM. Safety enhancing features used in EXAM. Additional data-safety-enhancing features were assessed by only sharing a certain percentage of weight updates with the largest magnitudes before sending them to the server after each round of learning52. We show that by using partial weight updates during FL, models can be trained that reach a performance comparable to training while sharing the full information. This differential privacy technique decreases the risk for model inversion or reconstruction of the training image data through gradient interception.



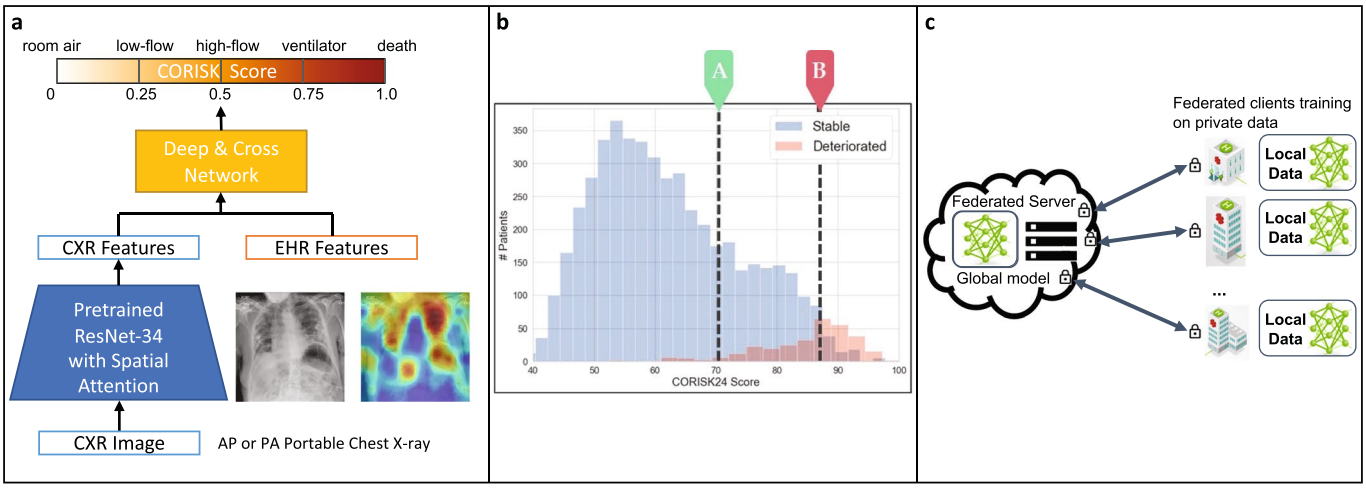
Extended Data Fig. 6 | Characteristics of EMR data used in EXAM. Characteristics of EMR data used in EXAM. Min. and max. values (asterisks) and mean and standard deviation (length of bars) for each EMR feature used as an input to the model. n specifies the number of sites that had this particular feature available. Missing values were imputed using a MissedForest algorithm.



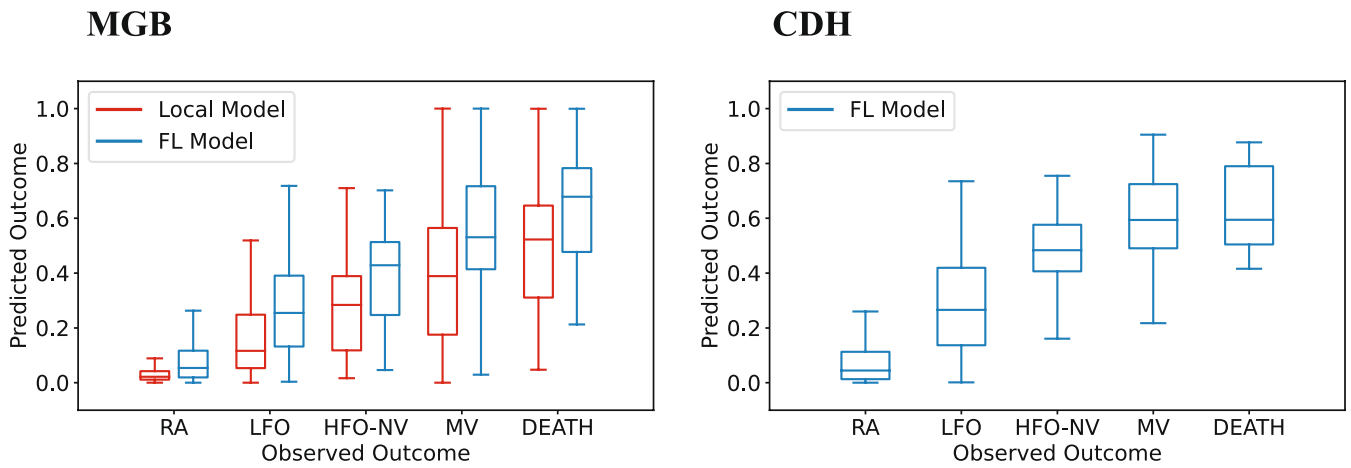
Extended Data Fig. 7 | Distribution of oxygen treatments between EXAM sites. Distribution of oxygen treatments between EXAM sites. The boxplots show the quartiles of the minimum, the maximum, the sample median, and the first and third quartiles (excluding outliers) of the oxygen treatments applied at different sites at time of Emergency Department admission and after 24 and 72- hour periods. The types of oxygen treatments administered are 'room air', 'low-flow oxygen', 'high-flow oxygen (non-invasive)', and 'ventilator'.



Extended Data Fig. 8 | Site variations in oxygen usage. Site variations in oxygen usage. Normalized distributions of oxygen devices at different time points, comparing the site with largest dataset size (site 1) and a site with unbalanced data, including mostly mild cases (site #16).



Extended Data Fig. 9 | Description of the EXAM Federated Learning study. Description of the EXAM Federated Learning study. **a**, Previously developed model, CDS, to predict a risk score that corresponds to respiratory outcomes in patients with SARS-COV-2. **b**, Histogram of CORISK results at MGB, with an illustration of how the score can be used for patient triage, in which 'A' is an example threshold for safe discharge that has 99.5% negative predictive value, and 'B' is an example threshold for Intensive Care Unit (ICU) admission that has 50.3% positive predictive value. For the purpose of the NPV calculation (threshold A), we defined the Model Inference to be Positive if it predicted oxygen need as LFO or above (COVID risk score ³0.25) and Negative if it predicted oxygen need as RA (<0.25). We defined the Disease to be Negative if the patient was discharged and not readmitted, and Positive if the patient was readmitted for treatment. For the purpose of PPV calculation (threshold B), we defined the Model Inference to be Positive if it predicted oxygen need as MV or above (³0.75) and Negative if it predicted oxygen need as HFO or less (<0.75). We defined the disease to be Positive if the patient required MV or if they died, and we defined the disease as Negative if the patient survived and did not require MV. The EXAM score can be used in the same way. **c**, Federated Learning using a client-server setup.



Extended Data Fig. 10 | Calibration Plots for the MGB data and the new independent dataset, CDH, used for model validation. Calibration Plots for the MGB data and the new independent dataset, CDH, used for model validation.

nature portfolio

Corresponding author(s): Mona Flores

Last updated by author(s): Aug 12, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Each site used different software in order to mine their medical records or data repositories for the required data. We do not have any visibility into what software they used to extract/prepare their data.
Data analysis	We used Client-server based federated learning (FL) using the FederatedAveraging algorithm (add number) as implemented in NVIDIA Clara Train v3.1 SDK. This code is freely available for use through NVIDIA Clara Imaging. https://developer.nvidia.com/clara-medical-imaging . For the model pipeline, we used the following: FL training software: Clara Train SDK v3.1.01 Model: clara_train_covid19_exam_ehr_xray v1 Preprocessing: pydicom==2.0.0 missingpy==0.2.0 (which includes the referenced 'MissForest' function) sklearn==0.0 imageio==2.9.0 numpy==1.19.1 scipy==1.5.2 Pillow==7.2.0 pandas==1.1.1 matplotlib==3.3.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The dataset from the 20 institutes that participated in this study remains under their custody. This data was used for training at each of the local sites and was not shared with any of the other participant institutions or with the Federated Server, and it is not publicly available. The dataset from the 3 independent test sites is maintained by Massachusetts General Brigham (MGB), and can be requested for access following appropriate procedures according to MGB IRB and policy (contact: Dr. Quanzheng Li).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The 20 client-sites prepared 16,148 cases (both positive and negative) for the purpose of training, validating, and testing the model. Each case included one CXR and the requisite data inputs taken from the patient's medical record. Based on our experience with training CORISK (the MGH local model that inspired the EXAM federated model), we expected that over 5,000 cases (roughly) should bring a model to performance, and additional data diversity would contribute to model robustness. We did not have a way to assess the amount nor diversity of the data needed to achieve robustness as there is no benchmark nor reliable methods to determine that in a non-empiric way.
Data exclusions	No data was excluded from the model training. In Fig. 3, Federated Learning vs. local training performance, we show the performance for 18 of 20 clients here as client 12 had only outcomes for 72 hours (see Extended Data Fig. 7) and client 14 only cases with room air treatment, resulting in the evaluation metric (avg. AUC) being not applicable (see Methods). This decision was established preemptively.
Replication	Each experiment was replicated 3 times, apart from the experiment evaluating efficacy of differential privacy. All attempts were successful.
Randomization	Data was randomized at the sites into train, validation and testing data.
Blinding	The main motivation for conducting federated learning is that the training of a deep learning model can occur without participants sharing or transferring data. This means that everyone was blinded to the data collection/allocation. The analysis took place after all training was complete, and incorporated all of the available information.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Patient data inclusion criteria were: 1. patient presented to the hospital's ED or equivalent, 2. patient had a RT-PCR test done anytime between presentation to the ED and discharge from the hospital, 3. patient had a CXR in the ED, 4. Patient's record had at least 5 of the EMR values detailed in Extended Data Table 1, all obtained in the ED, and the relevant outcomes captured during the hospitalization. Of note, The CXR, lab values, and vitals used were the first available captured during the visit to the ED. The model did not incorporate any CXR, lab values, or vitals acquired after leaving the ED.
Recruitment	This was a retrospective data study. Each institute completed their own review process in order to include their retrospective data in the study.
Ethics oversight	All procedures were conducted in accordance with principles for human experimentation as defined in the Declaration of Helsinki and International Conference on Harmonization Good Clinical Practice guidelines and approved by the relevant institutional review boards at the following validation sites: Cooley Dickinson Hospital (CDH), Martha's Vineyard Hospital (MVH), Nantucket Cottage Hospital (NCH), and at the following training sites: Mass Gen Brigham (MGB), Mass General Hospital (MGH), Brigham and Women's Hospital, Newton-Wellesley Hospital, North Shore Medical Center, Faulkner Hospital (all eight of these hospitals were covered under MGB's ethics board reference # 2020P002673 and informed consent was waived by the IRB). Similarly, the participation of the remaining sites was approved by their respective relevant institutional review processes: Children's National Hospital in Washington, D.C. (00014310, IRB Certified Exempt), NIHR Cambridge Biomedical Research Centre (20/SW/0140, Informed consent waived), The Self-Defense Forces Central Hospital in Tokyo (02-014, Informed consent waived), National Taiwan University MeDA Lab and MAHC and Taiwan National Health Insurance Administration (202108026W, Informed consent waived), Tri-Service General Hospital in Taiwan (B202105136, Informed consent waived); Kyungpook National University Hospital in South Korea (KNUH 2020-05-022, Informed consent waived), Faculty of Medicine, Chulalongkorn University in Thailand (490/63, 291/63, Informed consent waived), Diagnosticos da America SA in Brazil (26118819.3.0000.5505, Informed consent waived), University of California, San Francisco (20-30447, Informed consent waived), VA San Diego (H200086, IRB Certified Exempt), University of Toronto (20-0162-C, Informed consent waived), National Institutes of Health in Bethesda, Maryland (12-CC-0075, Informed consent waived), University of Wisconsin-Madison School of Medicine and Public Health (2016-0418, Informed consent waived), Memorial Sloan Kettering Cancer Center in New York (20-194, Informed consent waived), and Mount Sinai Health System in New York (IRB-20-03271, Informed consent waived).

Note that full information on the approval of the study protocol must also be provided in the manuscript.