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

Journal
Nature Medicine, 27(10)

ISSN
1078-8956

Authors
Dayan, Ittai
Roth, Holger R
Zhong, Aoxiao
et al.

Publication Date
2021-10-01

DOI
10.1038/s41591-021-01506-3

Peer reviewed

Published in final edited form as:

Nat Med. 2021 October ; 27(10): 1735–1743. doi:10.1038/s41591-021-01506-3.

γCorrespondence and requests for materials should be addressed to Mona G. Flores, MD. mflores@nvidia.com.
*These authors contributed equally: Ittai Dayan, Holger Roth, Aoxiao Zhong, Fiona J Gilbert, Quanzheng Li, Mona G. Flores.

Contributions

Ittai Dayan and Mona G. Flores contributed to the acquisition of the data, study support, drafting and revising the manuscript, study design, study concept, and analysis and interpretation of the data; Holger Roth, Aoxiao Zhong and Quanzheng Li, contributed to the acquisition of the data, study support, drafting and revising the manuscript, study design, and analysis and interpretation of the data; Fiona J Gilbert contributed to the acquisition of the data, study support, drafting and revising the manuscript; Jiahui Guan contributed to the support of the study, drafting and revising the manuscript, and analysis and interpretation of the data; Varun Buch contributed to the acquisition of the data, study support and study design; Daguang Xu contributed to the acquisition of the data, study support, drafting and revising the manuscript, and analysis and interpretation of the data; Anthony Beardsworth Costa, Bradford J. Wood, John W. Garrett and Krishna Juluru contributed to the acquisition of the data and drafting, and revising the manuscript; Nicola Rieke, contributed to the support of the study, and drafting and revising the manuscript; Ahmed Harouni, Anas Z Abidin, Andrew Liu, CK Lee, Colleen Ruan, Eddie Huang, Griffin Lacey, Jesse Tetreault, Kristopher Kersten, Pedro Mario Cruz e Silva, Abood Quraini, Andrew Feng, Colin Compas, Deepeksha Bhatia, Isaac Yang, Mohammad Adil and Yuhong Wen contributed to the support of the study; Amilcare Gentili, Chien-Sung Tsai, Chih-Hung Wang, Chun-Nan Hsu, Dufan Wu, Felipe Campos Kitamura,Gustavo César de Antônio Corradi, Gustavo Nino, Hao-Hsin Shin, Hirofumi Obinata, Hui Ren, Jason C. Crane, Joshua D Kaggie, Jung Gil Park, Keith Dreyer, Marcio Aloisio Bezerra Cavalcanti Rockenbach, Marius George Linguraru, Masoom A. Haider, Meena AbdelMaseeh, Pablo F. Damasceno, Pochuan Wang, Sheng Xu, Shuichi Kawano, Sira Sriswasdi, Soo Young Park, Thomas M. Grist, Watsamon Jantarabenjakul, Weichung Wang, Won Young Tak, Xiang Li, Xihong Lin, Young Joon Kwon, Andrew N Priest, Baris Turkbey, Benjamin Glicksberg, Bernardo Bizzo, Byung Seok Kim, Carlos Tor-Díez, Chia-Cheng Lee, Chia-Jung Hsu, Chin Lin, Chiu-Ling Lai, Christopher P. Hess, Eric K Oermann, Evan Leibovitz, Hisashi Sasaki, Hitoshi Mori, Jae Ho Sohn, Krishna Nand Keshava Murthy, Li-Chen Fu, Matheus Ribeiro Furtado de Mendonça, Mike Fralick, Min Kyu Kang, Natalie Gangai, Peerapon Vateekul, Pierre Elnajjar, Sarah Hickman, Sharmila Majumdar, Shelley L. McLeod, Sheridan Reed, Stefan Gräf, Stephanie Harmon, Tatsuya Kodama, Thanyawee Puthanakit, Tony Mazzulli, Vitor Lima de Lavor, Yothin Rakvongthai and Yu Rim Lee contributed to the acquisition of the data.

Competing interests

Financial 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., 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. declared ownership of NVIDIA Stock.

I.D. is presently an officer and shareholder of a company, Rhino HealthTech Inc., that provides systems for distributed computation, that can among other things, be used to complete federated learning tasks. He was not employed by this company during the execution of the EXAM study.

The remaining authors declare no competing interests.

Non-financial competing interests

C.H. declared Research travel, Siemens Healthineers AG; Conference Travel, EUROKONGRESS; GmBH; and Personal fees (Consultant, GE Healthcare LLC; DSMB Member, Focused Ultrasound Foundation).

F.J.G declared research collaborations with Merantix, Screen-Point, Lunit and Volpara, GE Healthcare and undertakes paid consultancy for Kheiron and Alphabet.

M.L. declared that he is the co-founder of PediaMetrix Inc. and is on the Board of the SIPAIM Foundation

S.E.H declared research collaborations with Merantix, Screen-Point, Lunit and Volpara.

B.J.W and S.X. declared 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 Grants Z1A CL040015, 1ZIDBC011242. Work supported by the NIH Intramural Targeted Anti-COVID-19 (ITAC) Program, funded by the National Institute of Allergy and Infectious Diseases. NIH may have intellectual property in the field.

The remaining authors declare no competing interests.

Editor summary:

Federated learning, a method for training artificial intelligence algorithms that protects data privacy, was used to predict future oxygen requirements of symptomatic COVID-19 patients using data from 20 different institutes across the globe.

Editor recognition statement:

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.

Reviewer recognition statement:

Nature Medicine thanks Fei Wang, Nikos Paragios and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

A full list of authors and affiliations appears at the end of the article.

Abstract

Federated learning (FL) is a method for training artificial intelligence (AI) models with data from multiple sources while maintaining the anonymity of the data, thus removing many barriers to data sharing. Here we use data from 20 institutes across the globe to train a FL model, called "EXAM" (EMR CXR AI Model), that predicts future oxygen requirements of symptomatic COVID-19 patients using inputs of vital signs, laboratory data, and chest X-rays. EXAM achieved an average area under the curve (AUC) greater than 0.92 for predicting outcomes at 24 and 72h from the time of initial presentation to the ER, and it provided a 16% improvement of the average AUC measured across all participating sites, and an average increase in generalizability of 38% when compared to models trained at a single site using that site's data. For predicting mechanical ventilation (MV) treatment or death at 24 h in the future at the largest independent test site, EXAM achieved a sensitivity of 0.950 and a 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 predicting clinical outcomes in COVID-19 patients, setting the stage for broader use of FL in healthcare.

The scientific, academic, medical and data science communities have come together in the face of the pandemic crisis in order 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 artificial intelligence (AI) disciplines have always fostered open and collaborative approaches, embracing concepts such as open-source software, reproducible research, data repositories, and making anonymized datasets publicly available[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 tech companies into the healthcare data world[15–17].

A concrete example of these types of collaborations is our previous work on an AI-based SARS-COV-2 Clinical Decision Support (CDS) model. The 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 lab values that were shown in previous publications to be predictive of COVID-19 patient outcomes[18–21]. CXR was selected as the imaging input because it is widely available and commonly indicated by guidelines such as provided by ACR[22], Fleischner Society[23], the WHO[24], national thoracic societies[25], national health ministry COVID handbooks and radiology societies throughout the world[26]. The output of the CDS model was a score, termed 'CORISK'[27] that corresponded to oxygen support requirements, and that could aid in triaging patients by front-line 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 lacks in diversity[31,32], potentially resulting in over-fitting and lower generalizability. This can be mitigated by training with diverse data from multiple sites without centralising data[33] using methods such as Transfer Learning[34,35] or FL. FL is a method to train Artificial Intelligence (AI) models on disparate data sources, without data being transported or exposed outside its original location. While applicable to many industries, FL has recently been proposed for cross-institutional healthcare research[36].

FL 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 'un-trained' 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 the 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 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 heavily rely on operator-dependent assessments of patient condition (such as clinical impressions or reported symptoms). Rather, lab results, vital signs, an imaging study and a commonly captured demographic (ie, 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' (**E**MR Chest **X**-Ray **A**I **M**odel) 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 generalise better across healthcare systems.

Results

The EXAM Model Architecture

EXAM is based on the CDS model mentioned above[27]. In total, 20 features (19 from the EMR and a CXR) were used as input to the model. The outcome (i.e., “ground truth”) labels were assigned based on patient oxygen therapy after 24-hour and 72-hour periods from initial admission to the 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, 0.5, 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)/non-invasive ventilation (NIV), or 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 of 0 to 1, corresponding to each of the prediction windows (ie, 24-hours and 72 hours).

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

The model therefore fuses information from both the EMR features and CXR features, using a ResNet-34 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‘ section). The model output is a risk score, termed EXAM score, which is a continuous value in the range of 0 – 1 for each of the 24 hour and 72-hour predictions, corresponding to the labels described above.

Federating The Model

EXAM was trained using a cohort of 16,148 cases, making it into not only one of the first FL models for COVID-19, but also the largest and first multi-continent development projects in clinically-relevant AI (Fig. 1 a,b). Data between sites was not harmonized prior to being extracted, and in light of real-life clinical informatics circumstances, a meticulous harmonization of the data input was not conducted by the authors, as seen in Fig. 1 c,d.

We compared the 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 ($p<<1e-3$, Wilcoxon signed-rank test) of 16% (as defined by the 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 a 38% generalizability improvement (as defined by the 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 predicting 24h oxygen treatment compared to models trained only on a site’s own data (Fig. 2b). For the prediction results of 72h oxygen treatments, 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 (e.g., mostly mild cases of COVID-19) markedly benefited from the FL approach with a substantial improvement in prediction avg. AUC performance for the categories with only a few cases. This was evident at Client-site (#16), with 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 the 'local' model as well as models trained on larger datasets from 5 client-sites in the Boston area (Fig. 3b).

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

Validation At Independent Sites

Following the 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 of them in Massachusetts, USA. The model was not re-trained at these sites, and it was only used 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 data set, from CDH, are shown in Fig. 4. The operating point was set to discriminate between non-mechanical ventilation (MV) treatment and MV treatment (or death). The FL global trained model, EXAM, achieved an average AUC of 0.944 and 0.924 for 24/72h prediction tasks respectively (Table 2 b), which exceeded the average performance among sites used in training EXAM. For predicting MV treatment (or death) at 24 h, EXAM achieved a sensitivity of **0.950** and a specificity of **0.882 at CDH, and a** sensitivity of 1.000 and specificity of 0.934 at MVH. NCH did not have any cases with MV / death at 24h. As for 72h 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 72h, EXAM had a low false-negative rate of 7.1% at CDH. 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 case had a CXR with 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 adhere 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, an 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 investigating a partial weight-sharing scheme[50,51], we showed that models can reach a comparable performance even when only 25% of the weight updates are shared (Extended Data Fig. 5).

Discussion

To our knowledge, this study features the largest real-world healthcare FL study to date 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 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 on individual sites than any model that was trained only on local data. We believe that consistent improvement was achieved not only due to larger, but also a more diverse data set, the use of data inputs that can be standardized and avoidance of clinical impressions / reported symptoms. These factors played a significant 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, indicate that the benefit for client-sites with small datasets arising from participation in FL collaborations is substantial. This is likely 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 generalisable 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 underway 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 the publications, focused on diagnosis of COVID-19 or predicting 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 an RT-PCR test result, 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 (CT), a non-consensual diagnostic modality. The model's design was constrained to objective predictors, unlike many published studies that leveraged subjective clinical impressions. The data collected reflects varied incidence rates, and thus the 'population momentum' we encountered is more diverse. That implies that the algorithm can be useful for 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' data sets. Improvements to clinical information systems are needed in order 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.

As data was not centralized, it is 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-hour oxygen consumption in the patient. It is assumed that oxygen delivered to the patient equates with 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 significant statistics regarding failure causes, post-hoc, at most sites. However, we did study the 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 two categories: 1) low quality of input data, e.g. missing data or motion artifact in CXR; 2) out-of-distribution data, e.g. a very young patient.

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

A feature that would enhance these kinds of large-scale collaborations, is the ability to predict each client-site's contribution towards improving the global FL model. This will help in client-site selection, and prioritizing data acquisition and annotation efforts. The latter is especially important given the high costs and difficult logistics of these large consortia endeavours, and it will enable these endeavours to capture diversity rather than sheer quantity of data samples.

Future approaches may incorporate automated hyperparameter searching[55], neural architecture search[56], and other automated machine learning (AutoML)[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] in order to reduce the divergence between unbalanced client-sites. Future work might explore different types of normalization techniques in order to allow the training of AI models in FL more effectively when the clients' data is non-independent and identically distributed (non-IID).

Recent works on privacy attacks within the FL setting have raised concerns on data leakage during model training[59]. Meanwhile, the protection algorithms are still under-explored and constrained by multiple factors. While differential privacy algorithms[36,48,49] show good protection, they may weaken the model's performance.

The encryption algorithms, such as Homomorphic Encryption[60] shall maintain the performance but may significantly increase the message sizes 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 more validation, we envision the deployment of the EXAM model in the ED setting, as a way to evaluate risk on a per-patient and on a population level and for providing clinicians with an additional reference point when making the often-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.

Methods

Ethics Approval

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

Study Setting

The study included data from 20 institutions (Fig. 1a); Mass Gen Brigham (MGB) affiliated hospitals (Mass General Hospital (MGH), Brigham and Women's Hospital, Newton-Wellesley Hospital, North Shore Medical Center, Faulkner Hospital); Children's National Hospital in Washington, D.C.; 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. The dataset curation started in June 2020 and the last 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 end-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 of them in Massachusetts, USA. These three hospitals had different patient population characteristics than the training sites. The data used for the algorithm validation consisted of patients admitted to the emergency department at these sites between March 2020 to February 2021 and 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 purpose of training, validating, and testing the model (Fig. 1b). Medical data was pulled in relation to patients who satisfied the study inclusion criteria. Client-sites strived to include all the COVID positive cases they had 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 September 30, 2020. The sites also included other patients in the same period that had 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 intensities (pixel values) varied significantly among the sites due 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 distributions varied greatly between sites, as expected due 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 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 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.

Model input

In total, 21 EMR features were used as input to the model. The outcome (i.e., “ground truth”) labels were assigned based on patient requirements after 24-hour and 72-hour 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-hour and 72-hour periods. The difference in dataset distribution for the largest and the smallest client-sites can be seen in Extended Data Fig. 8. The number of positive COVID-19 cases, confirmed by a single RT-PCR test obtained anytime between presentation to the ED and discharge from the hospital, are listed in Supplemental Table 1. Each client-site was asked to randomly split its dataset into 3 parts, 70% for training, 10% for validation, and 20% for testing. For both the 24h and 72h outcome prediction models, the 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 the hospital with symptoms of COVID-19, with some experiencing rapid deterioration in respiratory function requiring different interventions in order to prevent or mitigate hypoxemia[62,63]. A critical decision made during the evaluation of a patient at the initial point of care or the ED, is whether the patient is likely to require more invasive or resource-limited countermeasures or interventions (such as mechanical ventilation 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 ICU[64]. In contrast, a patient who is at a 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 continued self-monitoring at home[65]. EXAM was developed to help triage these patients. Of note, the model is not approved by any regulatory agency at this time, and it should only be used for research purposes.

EXAM score

EXAM was trained using FL, and 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 hours and 72 hours after initial presentation to the ED. Extended Data Fig. 9b illustrates how CORISK and the EXAM score can be used for patient triage.

CXR images were pre-processed to select the Anterior Position image and exclude lateral view images, and then scale to a resolution of 224×224. As shown in Extended Data Fig. 9a, the model fuses information from both the EMR features and CXR features (based on a modified ResNet-34 with spatial attention[67] pre-trained on the CheXpert dataset)[68], and Deep & Cross network[69]. To converge these different data types, a 512-dimensional feature vector was extracted from each CXR image using a pre-trained ResNet-34, 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 of 0 – 1 for both the 24 hour and 72-hour 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[70] using the NVIDIA Clara Train SDK[71]. The average AUC for the classification tasks (≥ LFO, ≥ HFO/NIV, or ≥ MV) was calculated and used as the final evaluation metric, and normalization to zero-mean and unit variance. CXR images were pre-processed to select the right series and exclude lateral view images, then scaled to a resolution of 224×224[27].

Feature imputation & normalization

A MissForest algorithm[66] 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 the EMR-CXR data fusion using deep & cross network

To model the interactions of features from the EMR and CXR data on a case-level, a deep feature scheme was used, based on a Deep & Cross Network architecture[69]. 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, that 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.

Federated Learning Details

Arguably, the most established form of FL is implementing the *Federated Averaging* algorithm 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 the 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 SSL encryption and communication protocols. The server then sends an updated set of weights to each client site after the aggregation, and sites resume training locally. The server and client site iterate back and forth until the model converges (Extended Data Fig. 9c).

A pseudo-algorithm of FL is shown in the Supplemental Note #1. In our experiments, we set the number of federated rounds to be 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 (24h or 72h). 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 the 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 5e-5 and a stepwise learning rate decay with a factor 0.5 after every 40 epochs, which is important for the convergence of *FederatedAveraging*[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.

Due to the sensitivity of Batch Normalization (BN) layers[58] when dealing with different clients in a non-IID setting, we found the best model performance to occur when keeping the pre-trained ResNet34 with spatial attention[47] parameters fixed during FL training (i.e. using a learning rate of zero for those layers). The Deep & Cross network that combines image features with the EMR features does not contain BN layers and hence was not affected by BN's 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 were shared with the server. To be exact, the weight updates (aka. 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 it to 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 hr time point (see Fig. 2 and Extended Data Fig. 1). The null hypothesis was rejected with a one-sided p-value << 1e-3 in both cases. A Pearson's correlation was used to assess the generalizability (robustness of the avg. 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, df = 17 for the 24h model and r=0.62, p=0.003, df=16 for the 72h model). This indicates that dataset size alone is not the only factor in determining a model's robustness to unseen data.

To compare the ROC curves from local models trained in different sites, and the global FL one (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 = (AUC1-AUC2)/s, where s is the standard deviation of the bootstrap differences, and AUC1 and AUC2 are the corresponding bootstrapped AUC series. By comparing D with the normal distribution, we obtained the p-values illustrated in Supplemental Table 2.The results show that the null hypothesis was rejected with very small 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 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 5 different local sites are 245.7, 253.4, 342.3, 389.8, while the F-value 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 4 ground truth categories. Furthermore, the p-value of the ANOVA test on the FL model is <2e-16, indicating that the FL prediction scores are statistically significantly different among the different prediction classes.

Data availability

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. Data from the independent validation sites is maintained by CAMCA, and access can be requested by contacting Dr. Quanzheng Li. Based on determination by CAMCA, a data sharing review and amendment of IRB for research purpose can be conducted by MGB research administration and in accordance with MGB IRB and policy.

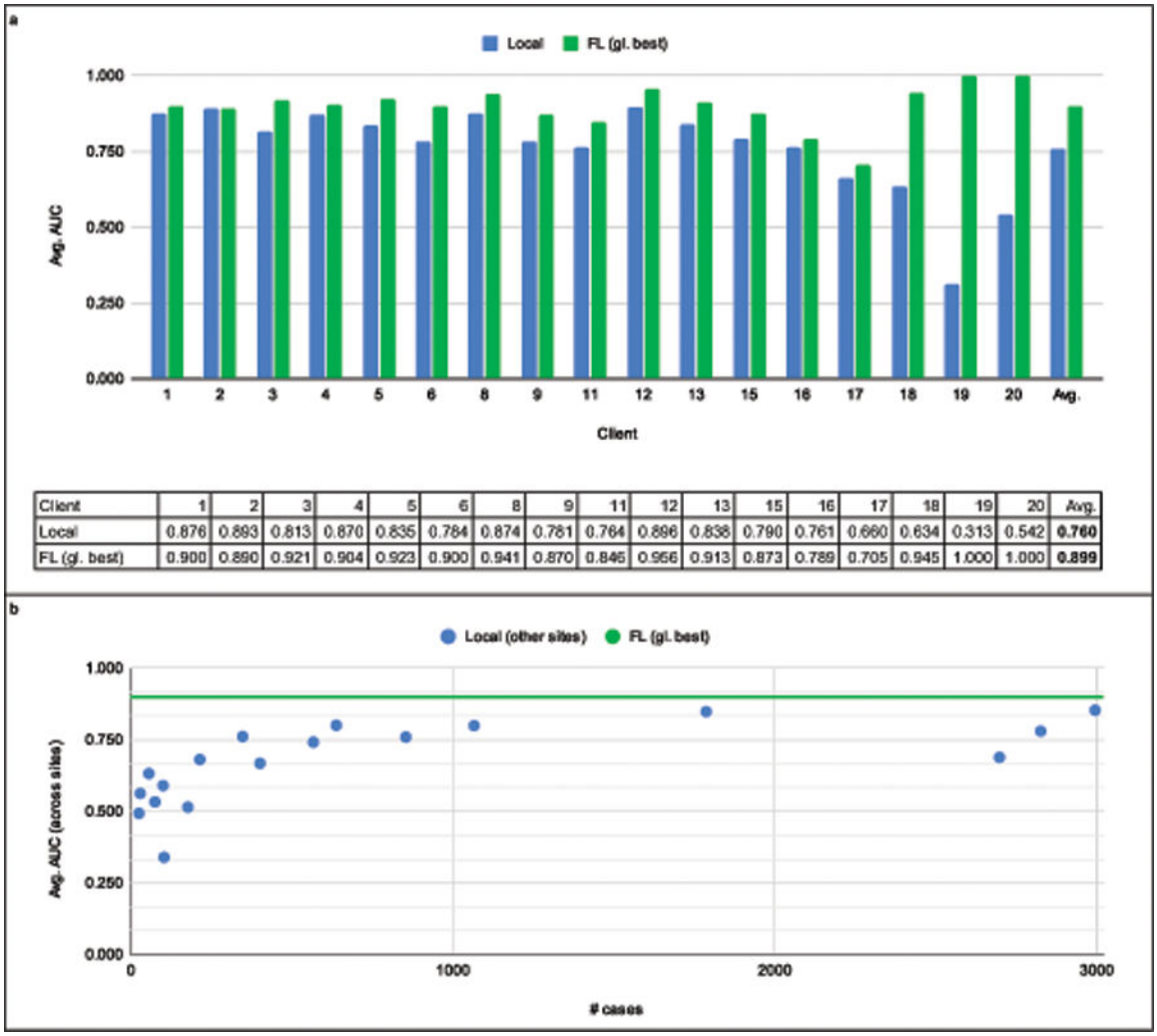
Code availability

All code and software used in this study are publicly available on NGC. To access, login as a Guest or create a profile, then enter one of the urls below. 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-Oclara_train_covid19_exam_ehr_xray_1.zip”.

The trained models, data preparation guidelines, code for training, validating testing 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

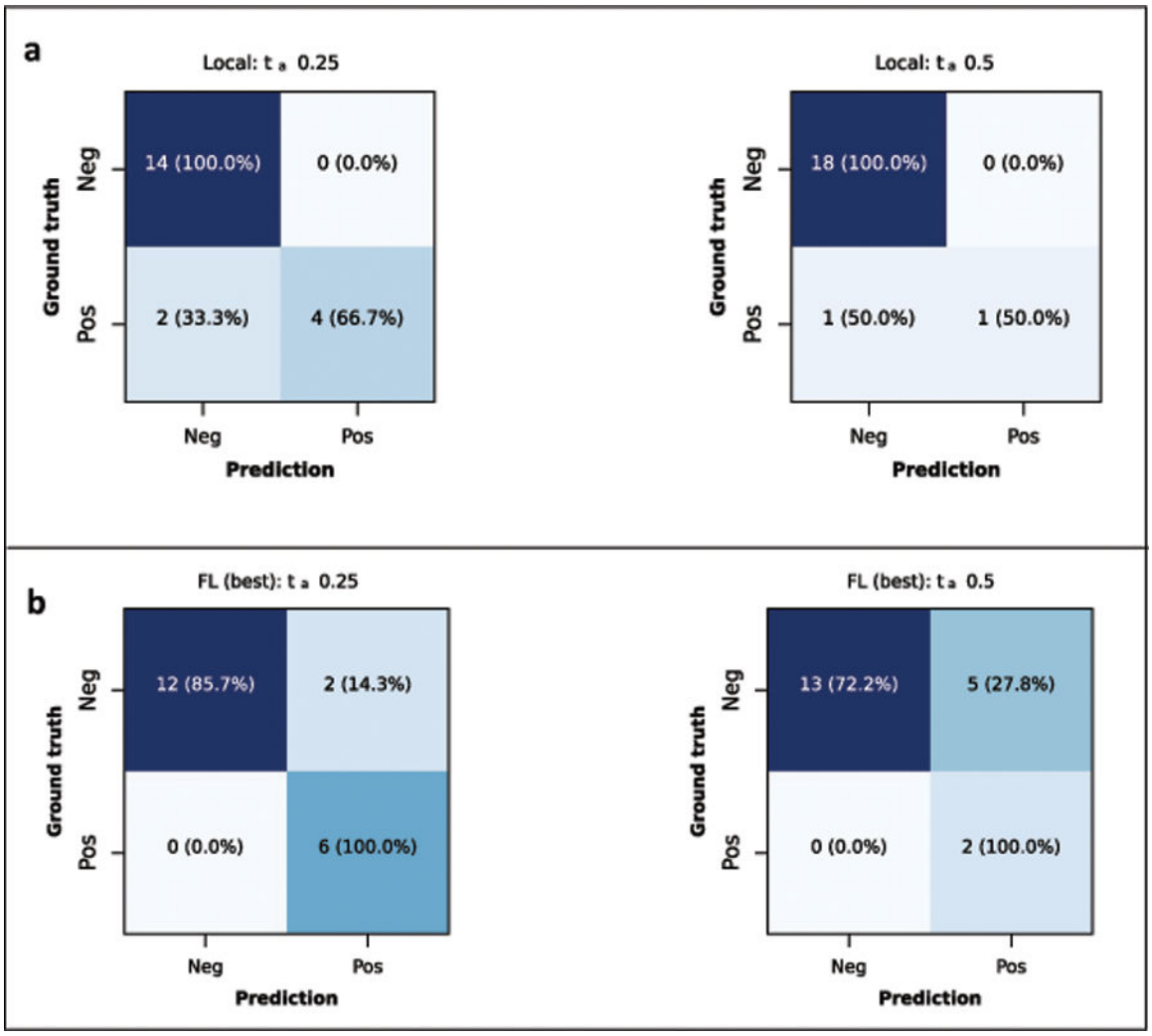
Extended Data



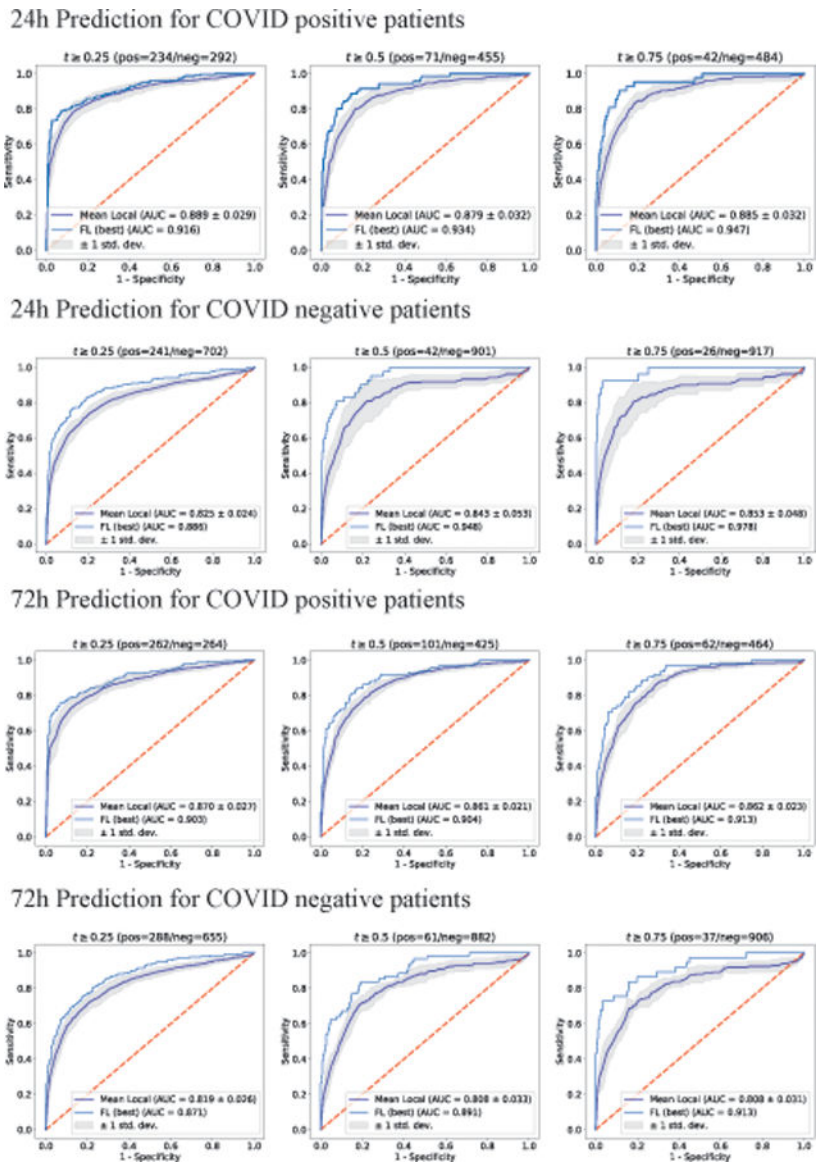
Client	1	2	3	4	5	6	8	9	11	12	13	15	16	17	18	19	20	Avg.
Local	0.876	0.893	0.813	0.870	0.835	0.784	0.874	0.781	0.764	0.896	0.838	0.790	0.761	0.660	0.634	0.313	0.542	**0.760**
FL (gl. best)	0.900	0.890	0.921	0.904	0.923	0.900	0.941	0.870	0.845	0.956	0.913	0.873	0.789	0.705	0.945	1.000	1.000	**0.899**

Extended Data Fig. 1. Test performance of models predicting 72h oxygen treatment trained on local data only versus the performance of the best global model available on the server.

Test performance of models predicting 72h 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).

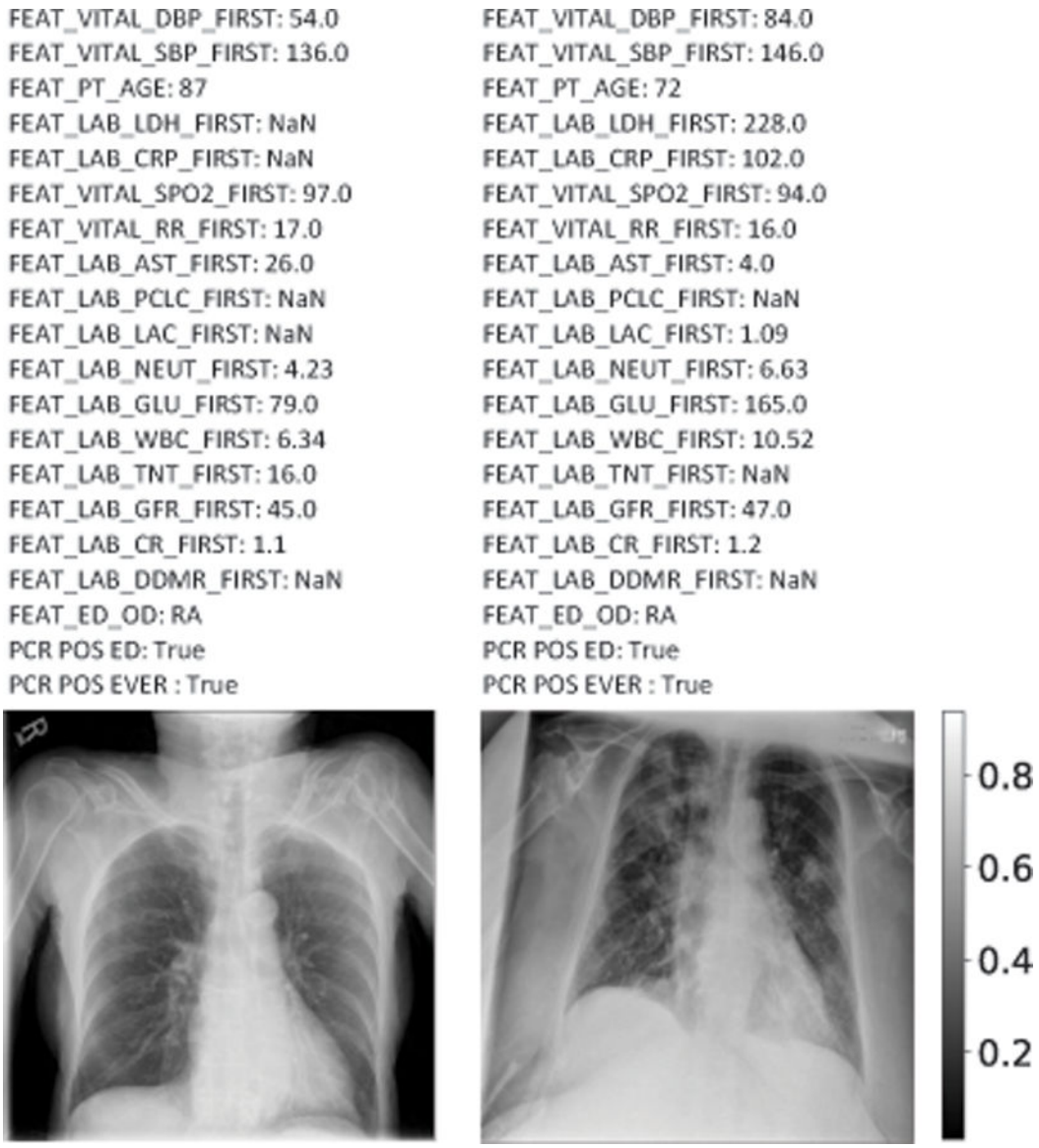


Extended Data Fig. 2. Confusion Matrices at a site with unbalanced data and mostly mild cases.
a, Confusion matrices on the test data at site 16 predicting oxygen treatment at 72h using the locally trained model. b, Confusion matrices on the test data at site 16 predicting oxygen treatment at 72h using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



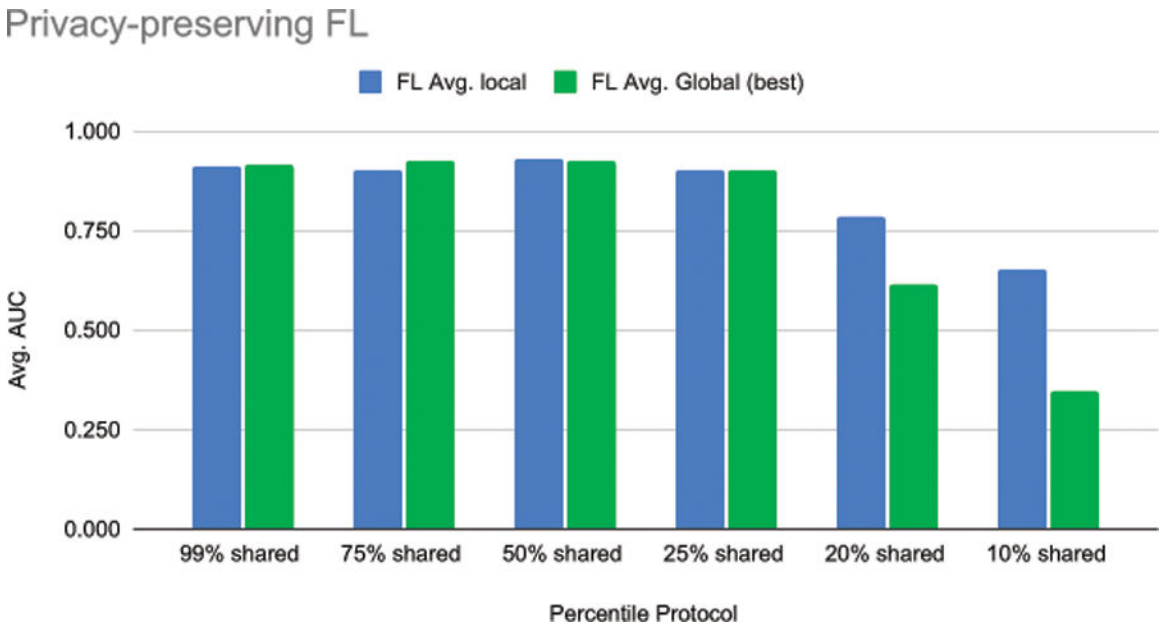
Extended Data Fig. 3. Effect of data set size on model performance.

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.



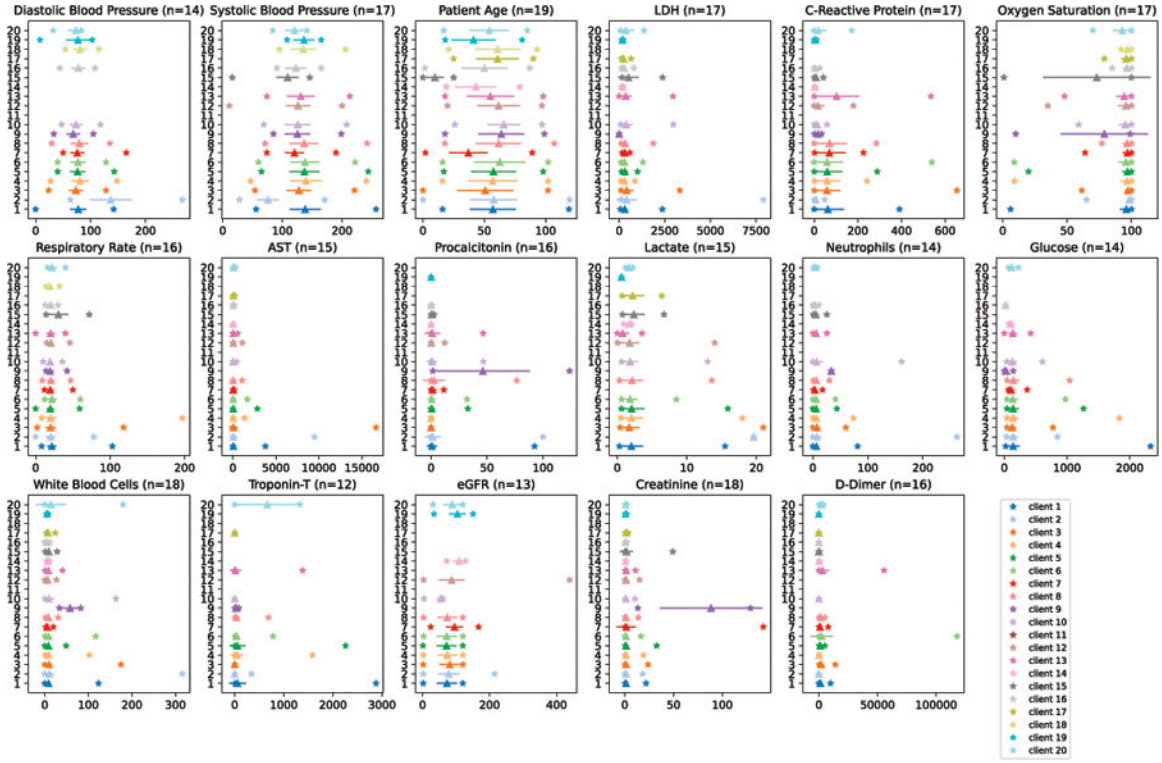
Extended Data Fig. 4. Failures cases at an independent test site.

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.



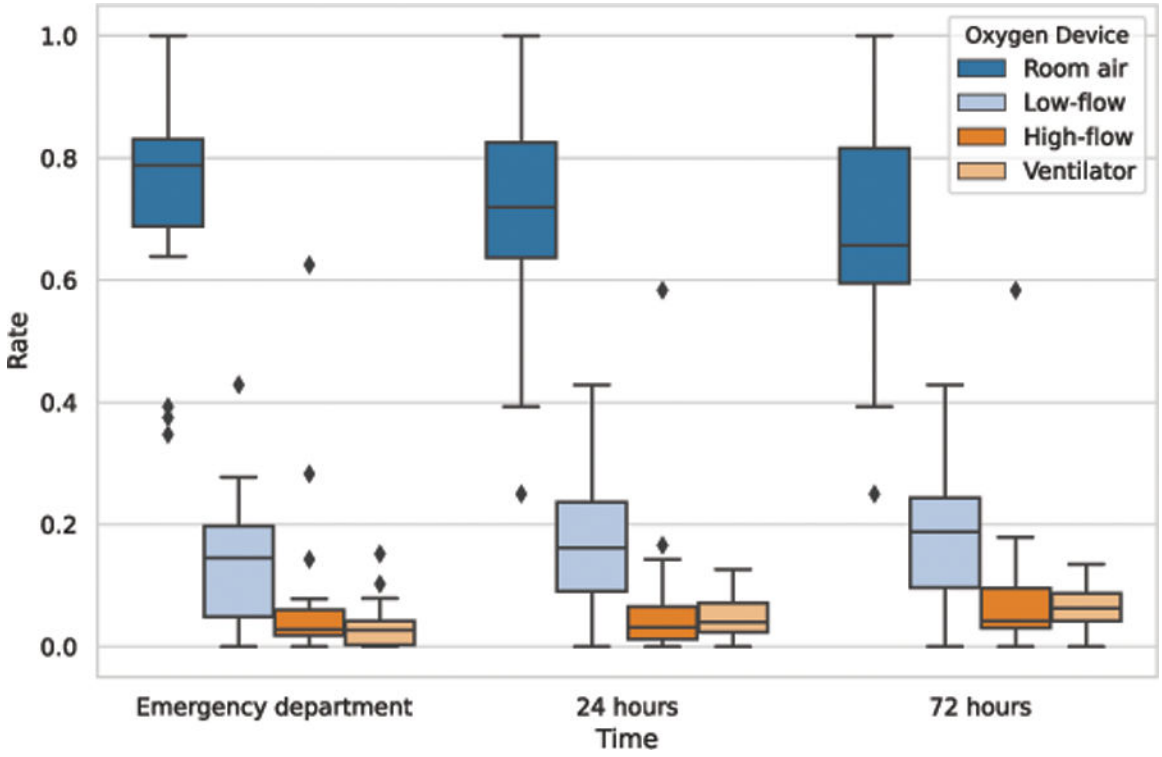
Extended Data Fig. 5. Safety enhancing features used in EXAM.

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.



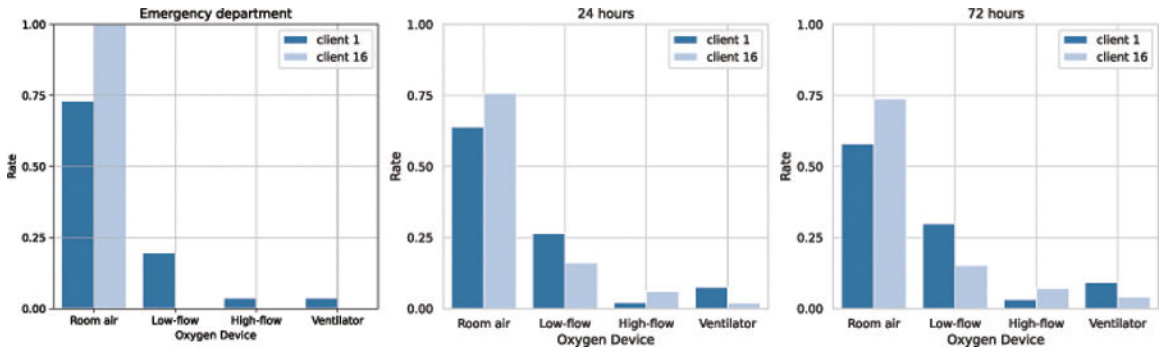
Extended Data Fig. 6. Characteristics of EMR data used in EXAM.

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. .



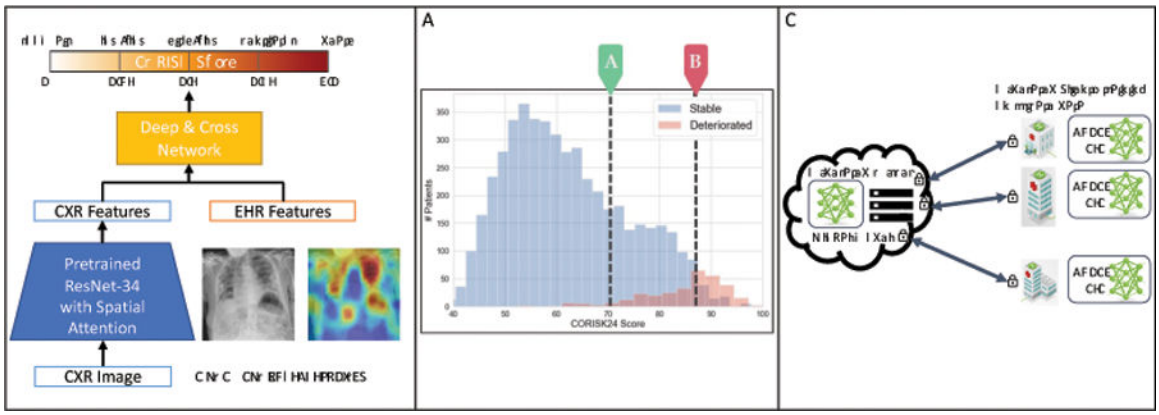
Extended Data Fig. 7. Distribution of oxygen treatments between EXAM sites.

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’.



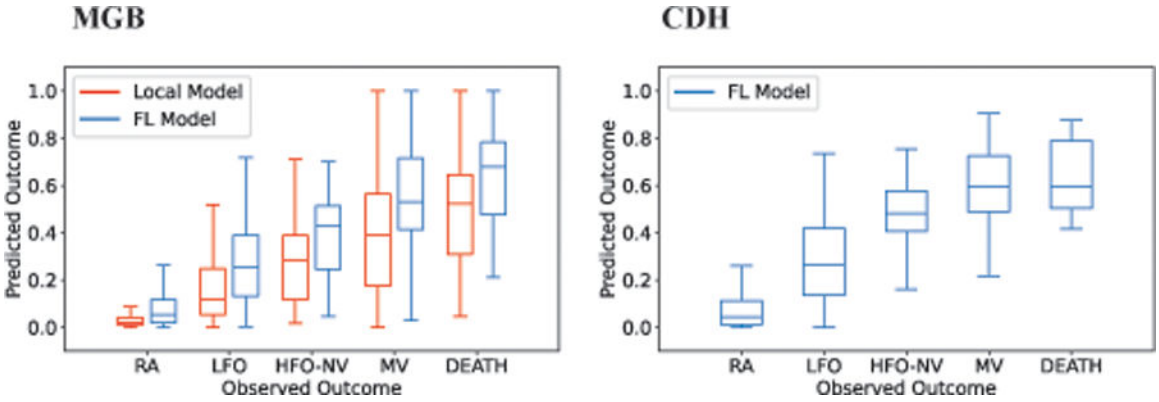
Extended Data Fig. 8. Site variations in oxygen usage.

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.

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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

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

Affiliations

[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]San Diego VA Health Care System, San Diego, CA, USA.

[5]Department of Neurosurgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[6]Radiology & Imaging Sciences / Clinical Center, National Institutes of Health, Bethesda, MD, USA.

[7]Division of Cardiovascular Surgery, Department of Surgery, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, R.O.C.

[8]Department of Otolaryngology-Head and Neck Surgery, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, R.O.C. and Graduate

Institute of Medical Sciences, National Defense Medical Center, Taipei, Taiwan, R.O.C.

[9]Center for Research in Biological Systems, University of California, San Diego, CA, USA.

[10]DasaInova, Diagnósticos da América SA (DASA), Brazil.

[11]Division of Pediatric Pulmonary and Sleep Medicine, Children's National Hospital, Washington, DC, USA.

[12]Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[13]Self-Defense Forces Central Hospital, Tokyo, Japan.

[14]Center for Intelligent Imaging, 2Department of Radiology and Biomedical Imaging, University of California, San Francisco, California, USA.

[15]Departments of Radiology and Medical Physics, The University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA.

[16]Department of Radiology, NIHR Cambridge Biomedical Resource Centre, University of Cambridge, Cambridge, UK.

[17]Department of Internal Medicine, Yeungnam University College of Medicine, Daegu, South Korea.

[18]Center for Clinical Data Science, Massachusetts General Brigham, Boston, MA, USA.

[19]Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, DC, USA.

[20]Joint Dept. of Medical Imaging, Sinai Health System, University of Toronto, Toronto, Canada and Lunenfeld-Tanenbaum Research Institute, Toronto, Canada.

[21]Lunenfeld-Tanenbaum Research Institute, Toronto, Canada.

[22]MeDA Lab and Institute of Applied Mathematical Sciences, National Taiwan University, Taipei, Taiwan.

[23]Research Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand, Center for Artificial Intelligence in Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.

[24]Department of Internal Medicine, School of Medicine, Kyungpook National University, Daegu, South Korea.

[25]Departments of Radiology, Medical Physics, and Biomedical Engineering, The University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA.

[26]Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand and Thai Red Cross Emerging Infectious Diseases Clinical Center, King Chulalongkorn Memorial Hospital, Bangkok, Thailand.

[27]Medical Review and Pharmaceutical Benefits Division, National Health Insurance Administration, Taipei. Taiwan.

[28]Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA.

[29]Department of Radiology, NIHR Cambridge Biomedical Resource Centre, Cambridge University Hospital. Cambridge, UK.

[30]Department of Radiology and Imaging Sciences, National Institutes of Health, Bethesda, MD, USA and National Cancer Institute, National Institutes of Health, Bethesda, MD, USA.

[31]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.

[32]Department of Internal Medicine, Catholic University of Daegu School of Medicine, Daegu, South Korea.

[33]Planning and Management Office, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, R.O.C.

[34]School of Medicine, National Defense Medical Center, Taipei, Taiwan, R.O.C. and School of Public Health, National Defense Medical Center, Taipei, Taiwan, R.O.C. and Graduate Institute of Life Sciences, National Defense Medical Center, Taipei, Taiwan, R.O.C.

[35]Department of Neurosurgery, NYU Grossman School of Medicine, New York, NY, USA.

[36]MOST/NTU All Vista Healthcare Center, Center for Artificial Intelligence and Advanced Robotics, National Taiwan University, Taipei, Taiwan.

[37]Division of General Internal Medicine and Geriatrics (Fralick), Sinai Health System, Toronto, Canada.

[38]Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand.

[39]Schwartz/Reisman Emergency Medicine Institute, Sinai Health, Toronto, ON, Canada and Department of Family and Community Medicine, University of Toronto, Toronto, ON, Canada.

[40]Department of Medicine and NIHR BioResource for Translational Research, NIHR Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK.

[41]National Cancer Institute, National Institutes of Health, Bethesda, MD, USA and Clinical Research Directorate, Frederick National Laboratory for Cancer, National Cancer Institute. Frederick, MD, USA.

[42]Department of Microbiology, Sinai Health/University Health Network, Toronto, Canada and Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto. Canada Public Health Ontario Laboratories, Toronto, Canada.

[43]Chulalongkorn University Biomedical Imaging Group and Division of Nuclear Medicine, Department of Radiology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.

[44]Departments of Radiology and Pediatrics, The George Washington University School of Medicine and Health Sciences, Washington, DC.

[45]School of Engineering and Applied Sciences, Harvard University, Boston, MA, US.

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 would like to acknowledge the following individuals for their support: James Brink MD, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA; Mannudeep Kalra MD, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA; Nir Neumark MD, MSc., Center for Clinical Data Science, Massachusetts General Brigham, Boston, MA; Thomas Schultz, Department of Radiology, Massachusetts General Hospital, Boston, MA; Ning Guo, Center for Advanced Medical Computing and Analysis, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA; Jayashree Kalpathy Cramer PhD, Director, QTIM lab at the Athinoula A. Martinos Center for Biomedical Imaging at MGH; Stuart Pomerant, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA; Giles Boland MD, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA; William Mayo-Smith MD, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

UCSF would like to acknowledge Peter B. Storey, Jed Chan and Jeff Block for implementing the UCSF FL client infrastructure and Wyatt Tellis, PhD 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.

Faculty of Medicine, Chulalongkorn University would like to acknowledge the Ratchadapisek Sompoch Endowment Fund RA (PO) 001/63 for the Collection and Management of COVID□ 19 Related Clinical Data and Biological Specimens for Research Task Force, Faculty of Medicine, Chulalongkorn University.

NIHR Cambridge Biomedical Research Centre would like to acknowledge that Andrew Priest is supported by the National Institute for Health Research (Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust).

National Taiwan University MeDA Lab and MAHC and Taiwan National Health Insurance Administration would like to acknowledge MOST Joint Research Center for AI Technology and All Vista Healthcare (AINTU) National Health Insurance Administration, Taiwan and Ministry of Science and Technology, Taiwan, and National Center for Theoretical Sciences Mathematics Division.

National Institutes of Health (NIH) would like to acknowledge that The National Institutes of Health (NIH) Medical Research Scholars Program is a public-private partnership supported jointly by the NIH and 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 National Institutes of Health.

Main References

1. Budd J et al. Digital technologies in the public-health response to COVID-19. Nat. Med 26, 1183–1192 (2020). [PubMed: 32770165]
2. Moorthy V, Henao Restrepo AM, Preziosi M-P & Swaminathan S Data sharing for novel coronavirus (COVID-19). Bull. World Health Organ 98, 150 (2020). [PubMed: 32132744]
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 (2020).
5. Candelon F, Reichert T, Duranton S, di Carlo RC & De Bondt M The Rise of the AI-Powered Company in the Postcrisis World (2020).

6. Chao H et al. Integrative analysis for COVID-19 patient outcome prediction. Medical image analysis 67, 101844 (2021). [PubMed: 33091743]
7. Zhu X et al. Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan. Medical image analysis 67, 101824 (2021). [PubMed: 33091741]
8. Yang D et al. Federated semi-supervised learning for Covid region segmentation in chest ct using multi-national data from china, italy, japan. arXiv 1–19 (2020) doi:10.1016/j.media.2021.101992.
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. Medical image analysis 65, 101794 (2020). [PubMed: 32781377]
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. Food and Drug Administration (FDA).Coronavirus Treatment Acceleration Program (CTAP). https://www.fda.gov/drugs/coronavirus-covid-19-drugs/coronavirus-treatment-acceleration-program-ctap (2020).
13. Gleeson P, Davison AP, Silver RA & Ascoli GA A Commitment to Open Source in Neuroscience. Neuron 96, 964–965 (2017). [PubMed: 29216458]
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). [PubMed: 29456894]
15. European Society of Radiology (ESR)., Neri E, de Souza N et al. What the radiologist should know about artificial intelligence – an ESR white paper. Insights Imaging 10, 44 (2019). 10.1186/s13244-019-0738-2 [PubMed: 30949865]
16. Pesapane F, Codari M & Sardanelli F Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. European Radiology Experimental 2, (2018).
17. Price WN 2nd & Cohen IG Privacy in the age of medical big data. Nat. Med 25, 37–43 (2019). [PubMed: 30617331]
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). [PubMed: 32396163]
19. Wynants L et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 369, m1328 (2020). [PubMed: 32265220]
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). [PubMed: 32306492]
21. Sands KE et al. Patient characteristics and admitting vital signs associated with coronavirus disease 2019 (COVID-19)-related mortality among patients admitted with noncritical illness. Infect. Control Hosp. Epidemiol 1–7 (2020) doi:10.1017/ice.2020.461.
22. ACR Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection | American College of Radiology https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection.
23. Rubin GD 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). [PubMed: 32255413]
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. American Thoracic Society. Diagnosis and Management of COVID-19 Disease 201, 15–19 (2020).
26. Redmond CE, Nicolaou S, Berger FH, Sheikh AM & Patlas MN Emergency Radiology During the COVID-19 Pandemic: The Canadian Association of Radiologists Recommendations for Practice. Canadian Association of Radiologists Journal 71, 425–430 (2020). [PubMed: 32468845]
27. Zhong A et al. Deep metric learning-based image retrieval system for chest radiograph and its clinical applications in COVID-19. Medical Image Analysis 70, 101993 (2021). [PubMed: 33711739]

28. Lyons C & Callaghan M The use of high-flow nasal oxygen in COVID-19. Anaesthesia 75, 843–847 (2020). [PubMed: 32246843]
29. Whittle JS, Pavlov I, Sacchetti AD, Atwood C & Rosenberg MS Respiratory support for adult patients with COVID-19. J Am Coll Emerg Physicians Open (2020) doi:10.1002/emp2.12071.
30. Ai J, Li Y, Zhou X & Zhang W COVID-19: treating and managing severe cases. Cell Res 30, 370–371 (2020). [PubMed: 32350393]
31. Esteva A et al. A guide to deep learning in healthcare. Nat. Med 25, 24–29 (2019). [PubMed: 30617335]
32. Cahan EM, Hernandez-Boussard T, Thadaney-Israni S & Rubin DL Putting the data before the algorithm in big data addressing personalized healthcare. npj Digital Medicine 2, 1–6 (2019). [PubMed: 31304351]
33. Thrall JH et al. Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. J. Am. Coll. Radiol 15, 504–508 (2018). [PubMed: 29402533]
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). [PubMed: 31932803]
35. Gao Y & Cui Y Deep transfer learning for reducing health care disparities arising from biomedical data inequality. Nat. Commun 11, 5131 (2020). [PubMed: 33046699]
36. Rieke N et al. The Future of Digital Health with Federated Learning (2020).
37. Yang Q, Liu Y, Chen T & Tong Y Federated Machine Learning: Concept and Applications (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 TS et al. Federated learning of predictive models from federated Electronic Health Records. Int. J. Med. Inform 112, 59–67 (2018). [PubMed: 29500022]
40. Roth HR et al. Federated Learning for Breast Density Classification: A Real-World Implementation: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings in Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning (eds. Albarqouni S et al.) vol. 12444 181–191 (Springer International Publishing, 2020).
41. Sheller MJ et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep 10, 12598 (2020). [PubMed: 32724046]
42. Remedios SW, Butman JA, Landman BA & Pham DL Federated Gradient Averaging for Multi-Site Training with Momentum-Based Optimizers. Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning 170–180 (2020) doi:10.1007/978-3-030-60548-3_17.
43. Xu Y et al. A collaborative online AI engine for CT-based COVID-19 diagnosis. medRxiv (2020) doi:10.1101/2020.05.10.20096073.
44. Raisaro JL et al. SCOR: A secure international informatics infrastructure to investigate COVID-19. Journal of the American Medical Informatics Association (2020) doi:10.1093/jamia/ocaa172.
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 Medical Informatics 9, (2021).
46. Nino G et al. Pediatric lung imaging features of COVID-19: A systematic review and meta-analysis. Pediatric Pulmonology 56, 252–263 (2021). [PubMed: 32926572]
47. Fredrikson M, Jha S & Ristenpart T Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security 1322–1333 (2015) doi:10.1145/2810103.2813677.
48. Zhu L, Liu Z & Han S Deep Leakage from Gradients. in Advances in Neural Information Processing Systems 32 (eds. Wallach H et al.) 14774–14784 (Curran Associates, Inc., 2019).
49. Kaissis GA, Makowski MR, Rückert D & Braren RF Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence vol. 2 305–311 (2020).
50. Li W et al. Privacy-Preserving Federated Brain Tumour Segmentation. Machine Learning in Medical Imaging 133–141 (2019) doi:10.1007/978-3-030-32692-0_16.

51. Shokri R & Shmatikov V Privacy-preserving deep learning 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton) (2015) doi:10.1109/allerton.2015.7447103.
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). [PubMed: 32679533]
53. Estiri H et al. Predicting COVID-19 mortality with electronic medical records. npj Digital Medicine 4, (2021).
54. Jiang G et al. Harmonization of detailed clinical models with clinical study data standards. Methods Inf. Med 54, 65–74 (2015). [PubMed: 25426730]
55. Yang D et al. Searching Learning Strategy with Reinforcement Learning for 3D Medical Image Segmentation. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 3–11 (2019) doi:10.1007/978-3-030-32245-8_1.
56. Elsken T, Metzen JH & Hutter F Neural Architecture Search: A Survey. arXiv [stat.ML] (2018).
57. Yao Q et al. Taking Human out of Learning Applications: A Survey on Automated Machine Learning. arXiv [cs.AI] (2018).
58. Ioffe S & Szegedy C Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv [cs.LG] (2015).
59. Kaufman S, Rosset S, Perlich C Leakage in Data Mining: Formulation, Detection, and Avoidance Proceedings of the 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 Proceedings of the 2020 USENIX Annual Technical Conference, ATC 2020 493–506 (2020).
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 JJ & Gattinoni L Management of COVID-19 Respiratory Distress. JAMA 323, 2329–2330 (2020). [PubMed: 32329799]
63. Cook TM 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). [PubMed: 32221970]
64. Galloway JB 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). [PubMed: 32479771]
65. Kilaru AS et al. Return Hospital Admissions Among 1419 COVID-19 Patients Discharged from Five U.S. Emergency Departments. Acad. Emerg. Med 27, 1039–1042 (2020). [PubMed: 32853423]
66. Stekhoven DJ & Bühlmann P MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118 (2012). [PubMed: 22039212]
67. He K, Zhang X, Ren S & Sun J Deep Residual Learning for Image Recognition 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) doi:10.1109/cvpr.2016.90.
68. Irvin J et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison Proceedings of the AAAI Conference on Artificial Intelligence vol. 33 590–597 (2019).
69. Wang R, Fu B, Fu G & Wang M Deep & Cross Network for Ad Click Predictions. Proceedings of the ADKDD'17 on ZZZ - ADKDD'17 (2017) doi:10.1145/3124749.3124754.
70. Chen Martın Abadi Jianmin, Chen Zhifeng, Davis Andy, Dean Jeffrey, Devin Matthieu, Ghemawat Sanjay, Irving Geoffrey, Isard Michael, Kudlur Manjunath, Levenberg Josh, Monga Rajat, Moore Sherry, Murray Derek G., Steiner Benoit, Tucker Paul, Vasudevan Vijay, Tensorflow PB: A system for large-scale machine learning 12th USENIX Symposium on Operating Systems Design and Implementation (2016) doi:10.1007/978-1-4842-6699-1_1.
71. NVIDIA Clara Imaging https://developer.nvidia.com/clara-medical-imaging (2020).
72. McMahan H, Moore E, Ramage D, Hampson S & y Arcas BA Communication-Efficient Learning of Deep Networks from Decentralized Data. in AISTATS (2017).

73. Hsieh K, Phanishayee A, Mutlu O & Gibbons PB The Non-IID Data Quagmire of Decentralized Machine Learning. arXiv [cs.LG] (2019).

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). [PubMed: 21414208]

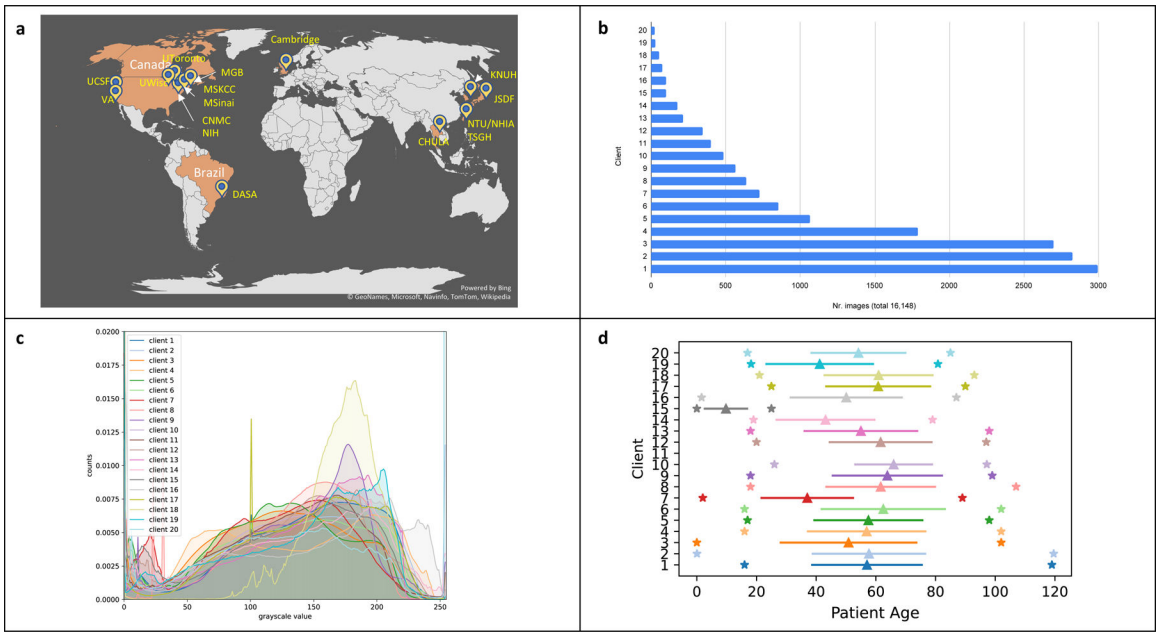
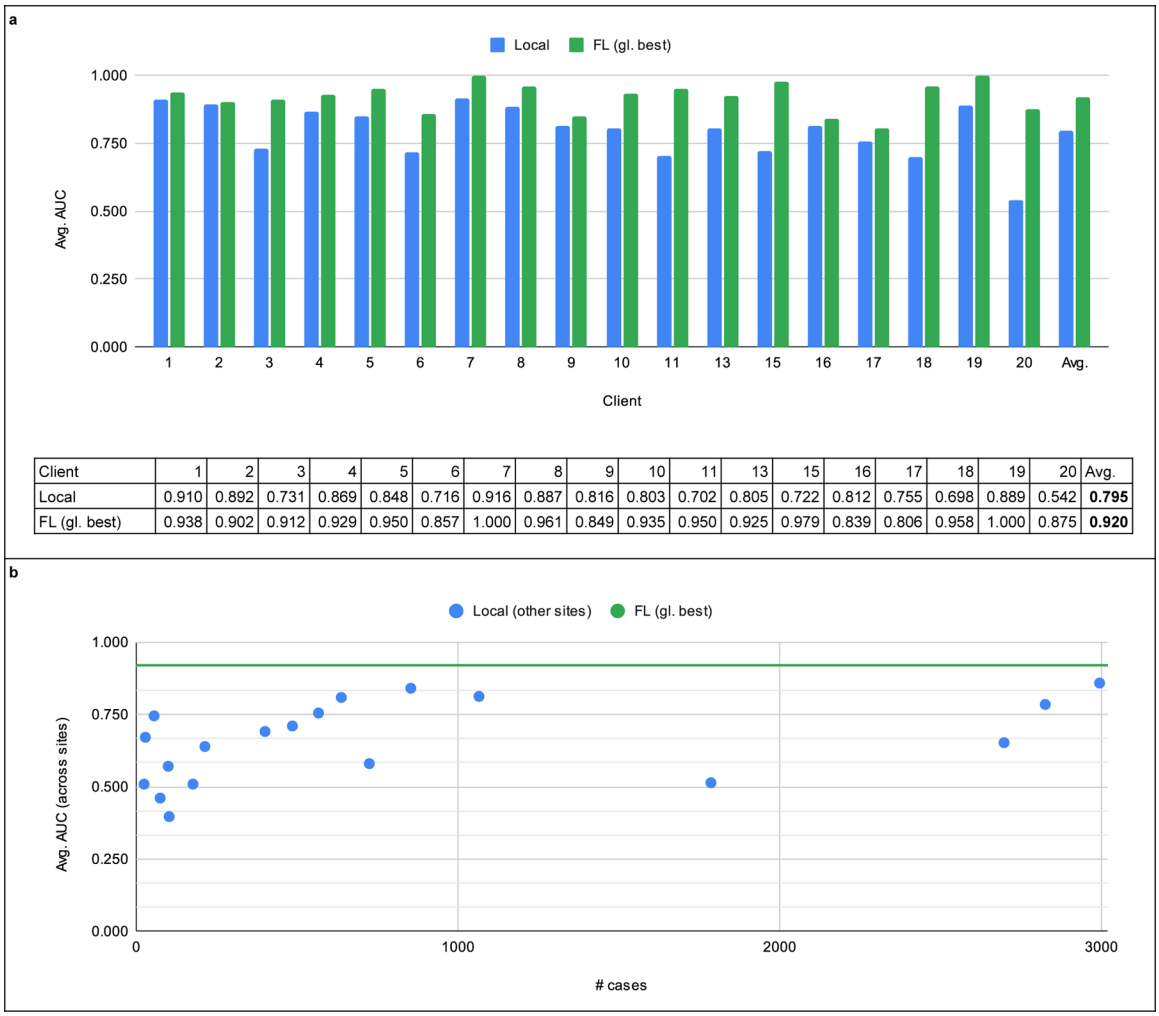


Fig. 1 |.

Data used in the EXAM FL study. a, World map indicating the 20 different client sites contributing to EXAM study. b, Number of cases contributed by each institution or site contributed (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 the minimum and maximum ages (asterisks), mean age (triangle) and standard deviation (horizontal bar). The number of samples of each client site is shown in Supplemental Table 1.



Client	1	2	3	4	5	6	7	8	9	10	11	13	15	16	17	18	19	20	Avg.
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 federated learning versus local models. a, Performance on each client's test set for predicting 24h oxygen treatment for models trained on local data only (Local) versus the performance of the best global model available on the server (FL (gl. best)). "Avg." stands for the 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 (# cases). The green horizontal line shows the generalizability performance of the best global model. The performance for 18 of 20 clients is shown, as client 12 had outcomes only for 72 hours (see Extended Data Fig. 1) and client 14 had cases only with room air treatment, such that the evaluation metric (avg. AUC) was not applicable in either of these cases (see Methods). Data for client 14 was also excluded from the computation of the average generalizability of the local models.

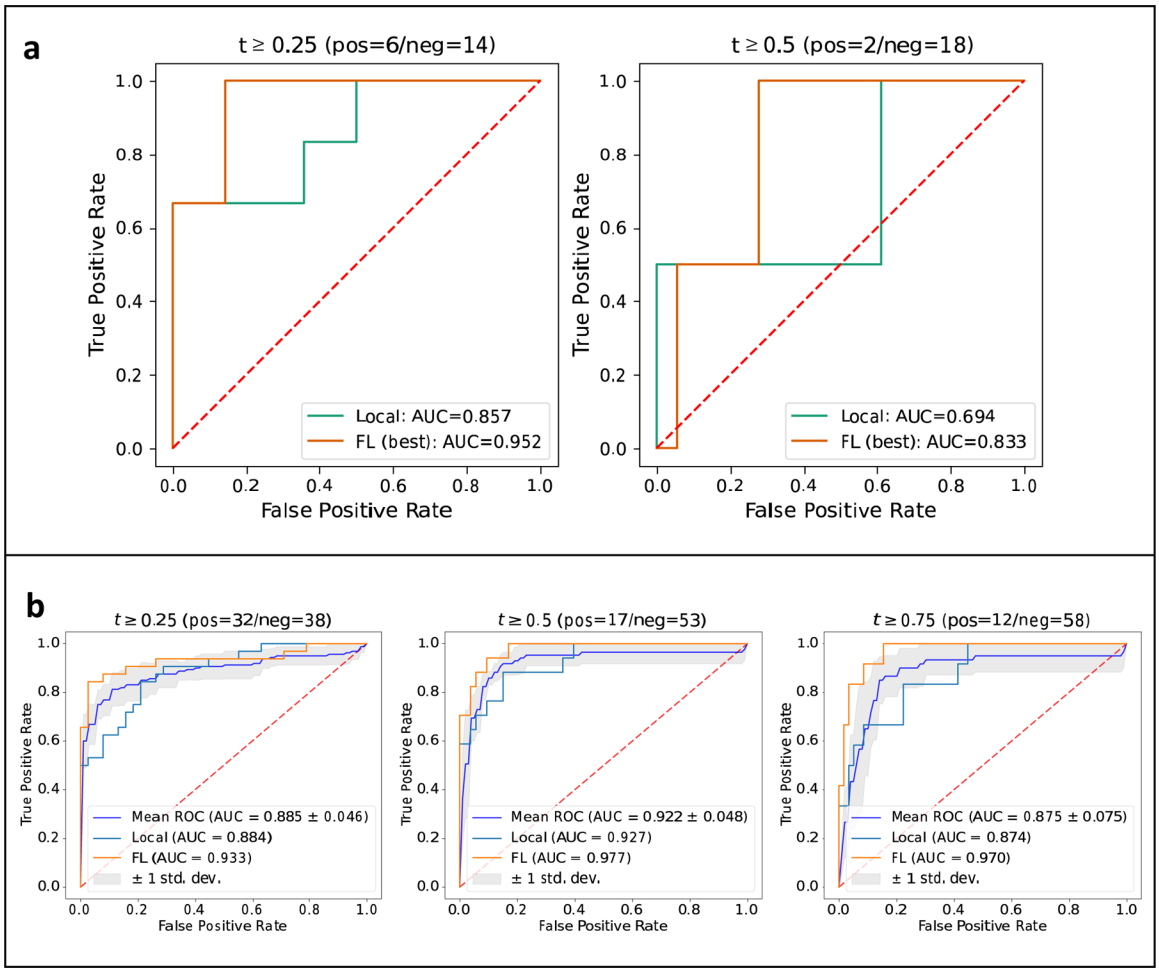


Fig. 3 |.

Comparison of federated learning-trained to locally-trained models. **a,** ROC at a site with unbalanced data and mostly mild cases (client-site #16)**. b,** ROC of the local model at client-site #12 (a small dataset), mean ROC of models trained on larger datasets corresponding to the 5 client-sites in the Boston area (#1, #4, #5, #6, #8), and ROC of the best global model to predict oxygen treatment at 72 h for different thresholds of the EXAM score (left, middle, right). The mean ROC is calculated based on 5 locally-trained models, and the gray-area shows the standard deviation of the ROCs. ROCs for three different cut-off values t of the EXAM risk score are shown. “Pos” and “Neg” stand for the number of positive and negative cases defined by this range of the EXAM score, respectively.

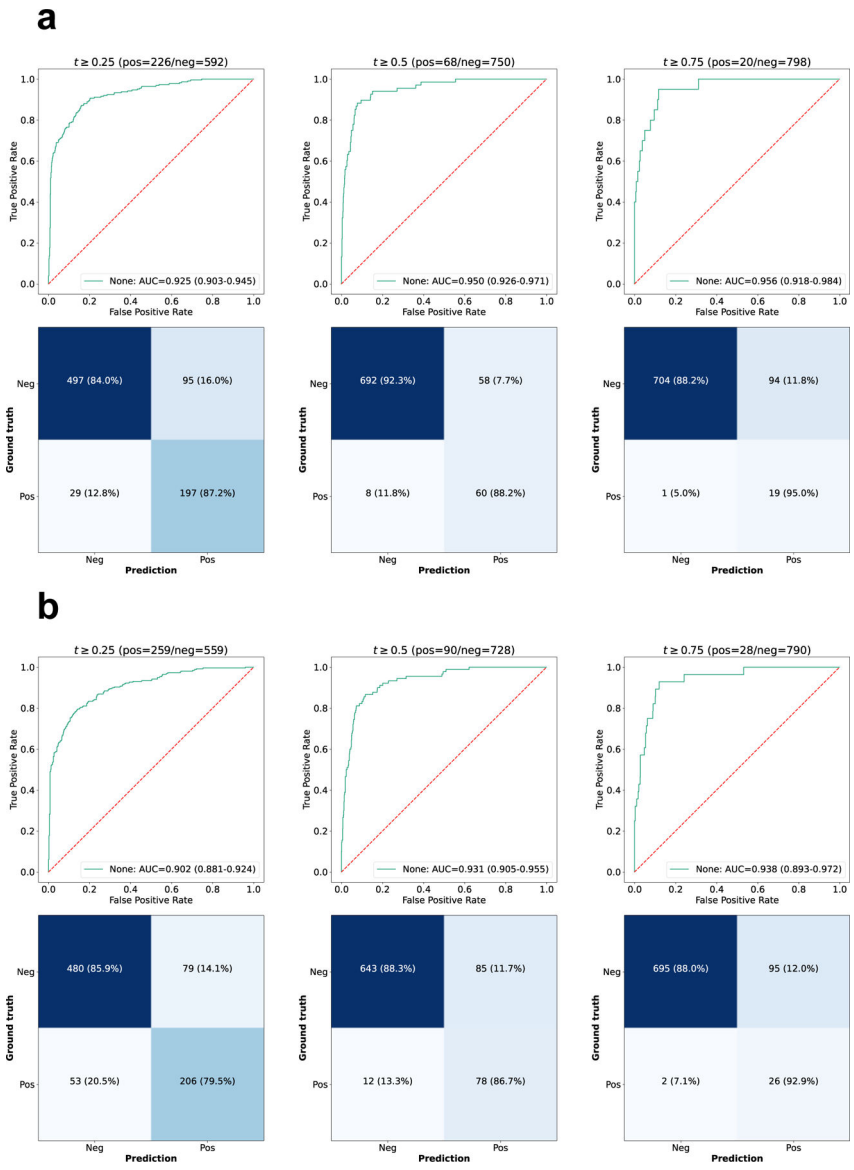


Fig. 4 |.

Performance of the best global model on the largest independent data set. **a,** Performance (ROC) (top) and confusion matrices (bottom) of the EXAM FL model on the CDH dataset for predicting oxygen treatment at 24 h. **b,** Performance (ROC) (top) and confusion matrices (bottom) of the EXAM FL model on the CDH dataset for predicting oxygen treatment at 72 h. ROCs for three different cut-off values ***t*** of the EXAM risk score are shown. “Pos” and “Neg” stand for the number of positive and negative cases defined by this range of the EXAM score, respectively.

Table 1 |

EMR (electronic medical record) data used in the EXAM study

Category	Subcategory	Component Name	Definition	Units	LOINC Code
Demographic	-	Patient Age	-	Years	30525–0
Imaging	Portable Chest X-Ray	-	AP or PA Portable Chest X-ray	-	36554–4
Lab Value	C-Reactive Protein	C Reactive Protein	Blood C-Reactive Protein Concentration	mg/L	1988–5
Lab Value	CBC (Complete Blood Count)	Neutrophils	Blood Absolute Neutrophils	10^9/L	751–8
Lab Value	CBC (Complete Blood Count)	White Blood Cells	Blood White Blood Cell Count	10^9/L	33256–9
Lab Value	D-Dimer	D-Dimer	Blood D-Dimer Concentration	ng/mL	7799–0
Lab Value	Lactate	Lactate	Blood Lactate Concentration	mmol/L	2524–7
Lab Value	LDH (Lactate Dehydrogenase)	LDH	Blood Lactate Dehydrogenase Concentration	U/L	2532–0
Lab Value	Metabolic Panel	Creatinine	Blood Creatinine Concentration	mg/dL	2160–0
Lab Value	Procalcitonin	Procalcitonin	Blood Procalcitonin Concentration	ng/mL	33959–8
Lab Value	Metabolic Panel	eGFR	Estimated Glomerular Filtration Rate	mL/min/ 1.73m2	69405–9
Lab Value	Troponin	Troponin-T	Blood Troponin Concentration	ng/ml	67151–1
Lab Value	Hepatic Panel	AST	Blood AST Concentration	IU/L	1920–8
Lab Value	Metabolic Panel	Glucose	Blood Glucose Concentration	mg/dL	2345–7
Vital Sign	-	Oxygen Saturation	Oxygen Saturation	%	59408–5
Vital Sign	-	Systolic Blood Pressure	Systolic Blood Pressure	mmHg	8480–6
Vital Sign	-	Diastolic Blood Pressure	Diastolic Blood Pressure	mmHg	8462–4
Vital Sign	-	Respiratory Rate	Respiratory Rate	breaths per minute	9279–1
Vital Sign		COVID PCR test	PCR for RNA [not used as input to model]		95425–5
Vital Sign	Oxygen Device used at Emergency Department (ED)	Oxygen Device	Ventilation, High-flow/NIV, Low-flow, Room Air	-	41925–9
Outcome	24Hr Oxygen Device	Oxygen Device	Ventilation, High-flow/NIV, Low-flow, Room Air	-	41925–9
Outcome	72Hr Oxygen Device	Oxygen Device	Ventilation, High-flow/NIV, Low-flow, Room Air	-	41925–9
Outcome	Death	-	-	-	-
Outcome	Time of Death	-	-	Hours	-

Table 2 |
Performance of EXAM on independent data sets.

a, Breakdown of patients by level of oxygen needed across the 3 independent datasets, CDH, MVH and NCH. b, AUC for predicting the level of oxygen needed at 24 and 72 h for the 3 independent datasets (with 95% confidence intervals). The AUC for the NCH dataset for MV at 24 h could not be calculated as there were no mechanically ventilated patients. AUC - area under the curve, RA- room air, LFO –low flow oxygen, HFO-NV – high flow oxygen, no mechanical ventilation, MV – mechanical ventilation, #Cases – number of patients included in the data set, # Pos. Cases – number of patients with confirmed COVID-19 infection included in the data set

a

Site	# Cases	# Pos. Cases	Prediction Interval	# Patients at each level of Oxygen Needed			
				RA	LFO	HFO-NV	MV&DEATH
CDH	840	244	24 hours	608	162	48	22
			72 hours	575	173	62	30
MVH	399	30	24 hours	356	36	3	4
			72 hours	351	39	3	6
NCH	264	29	24 hours	237	23	4	0
			72 hours	235	22	4	3

b

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