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

Citation

Permanent link
https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37377340

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Federated Learning used for predicting outcomes in SARS-COV-2 patients


1MGH Radiology and Harvard Medical School, Boston, MA, USA. NVIDIA, Santa Clara, CA, USA. Center for Advanced Medical Computing and Analysis, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.
2San Diego VA Health Care System, San Diego, CA, USA. Department of Neurosurgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA. Radiology & Imaging Sciences / Clinical Center, National Institutes of Health, Bethesda, MD, USA. Division of Cardiovascular Surgery, Department of Surgery, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, R.O.C.
3Department of Otologyngology-Head and Neck Surgery, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, R.O.C.
4Department of Radiology, University of California, San Diego, CA, USA.
5Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.
6Department of Internal Medicine, School of Medicine, Kyungpook National University, Daegu, South Korea.
7Departments of Radiology, Medical Physics, and Biomedical Engineering, The University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA.
8Department of Radiology, NIHR Cambridge Biomedical Resource Centre, University of Cambridge, Cambridge, UK.
9Department of Internal Medicine, Yeungnam University College of Medicine, Daegu, South Korea.
10Center for Clinical Data Science, Massachusetts General Brigham, Boston, MA, USA.
11Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, DC, USA. Joint Dept. of Medical Imaging, Sinai Health System, University of Toronto, Toronto, Canada and Lunenfeld-Tanenbaum Research Institute, Toronto, Canada. Lunenfeld-Tanenbaum Research Institute, Toronto, Canada. MeDa Lab and Institute of Applied Mathematical Sciences, National Taiwan University, Taipei, Taiwan.
12Center for Interventional Oncology, National Institutes of Health, Bethesda, MD, USA. Research Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.
13Center for Artificial Intelligence in Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.
14Department of Internal Medicine, School of Medicine, Chulalongkorn National University, Daegu, South Korea.
15Departments of Radiology, Medical Physics, and Biomedical Engineering, The University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA.
16Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.
17Thailand.
18Department of Medical Imaging, Medical Physics, and Biomedical Engineering, The University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA.
19Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.
20Medical Review and Pharmaceutical Benefits Division, National Health Insurance Administration, Taipei, Taiwan.
21Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA. Department of Radiology, NIHR Cambridge Biomedical Resource Centre, Cambridge University Hospital, Cambridge, UK. Department of Radiology and Imaging Sciences, National Institutes of Health, Bethesda, MD, USA and National Cancer Institute, National Institutes of Health, Bethesda, MD, USA.
22Hasso 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. Department of Internal Medicine, Catholic University of Daegu School of Medicine, Daegu, South Korea. Planning and Management Office, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, R.O.C.
23School of Medicine, National Defense Medical Center, Taipei, Taiwan, R.O.C. and Graduate Institute of Life Sciences, National Defense Medical Center, Taipei, Taiwan, R.O.C.
24Department of Neurosurgery, NYU Grossman School of Medicine, New York, NY, USA. MOST/NTU All Vista Healthcare Center, Center for Artificial Intelligence and Advanced Robotics, National Taiwan University, Taipei, Taiwan. Division of General Internal Medicine and Geriatrics (Fralick), Sinai Health System, Toronto, Canada.
25Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand.
26Schwartz/Reisman Emergency Medicine Institute, Sinai Health, Toronto, ON, Canada and Department of Family and Community Medicine, University of Toronto, Toronto, ON, Canada. Department of Medicine, NIHR Cambridge Biomedical Resource Centre, University of Cambridge, Cambridge, UK.
27National Cancer Institute, National Institutes of Health, Bethesda, MD, USA and Clinical Research Directorate, Frederick National Laboratory for Cancer, National Cancer Institute. Frederick, MD, USA. 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. Chulalongkorn University Biomedical Imaging Group and Division of Nuclear Medicine, Department of Radiology, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.
28These authors contributed equally: Ittai Dayan, Holger Roth, Aoxiao Zhong, Fiona J Gilbert, Quanzheng Li, Mona G. Flores. *e-mail: mfl ores@nvidia.com
Federated Learning (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, it has been recently proposed for cross-institutional healthcare research.

One approach to FL is to send an ‘un-trained’ model to other servers (‘nodes’), that conduct partial training tasks and send the results back to be merged in the central (‘federated’) server in an iterative process until training is complete.

In the field of Artificial Intelligence, training Deep Learning models requires a large amount of data. While healthcare data is abundant, it is generally siloed and not readily shared between different institutions due to privacy, regulatory and commercial concerns.

Here we show that FL facilitates training better models by unlocking learning from large datasets of sensitive healthcare data without centralizing or sharing the data.

At the onset of the SAR-COV-2 pandemic, we conducted a global Artificial Intelligence study to develop an AI model “EXAM” (EMR CXR AI Model), using FL, to assist in patient triage. We compared FL vs. local training in training a model to predict respiratory outcomes in patients with SARS-COV-2. The FL derived model had an AUC = 0.95 on real-world hospital data, outperforming the local training models by 16% on average (from 0.795 to 0.920 or 12.5 percentage points), and outperforming the local training models by 38% in terms of generalizability (from 0.667 to 0.920 or 25.3 percentage points).

EXAM, the largest healthcare FL project to-date, trained a model across diverse and distributed datasets. It is a powerful proof-of-concept of the feasibility of using FL for fast and collaborative development of needed AI models 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\textsuperscript{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\textsuperscript{3–10}. Clinical trial recruitment has been expedited and facilitated by national regulatory bodies and an international cooperative spirit\textsuperscript{11–13}. 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\textsuperscript{14,15}. 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\textsuperscript{16–19}.

A concrete example of these types of collaborations is our recent work on an AI-based SARS-COV-2 Clinical Decision Support (CDS) model. The CDS model, that was validated across multiple health systems’ data, predicts a risk score that could be used to support patient placement and risk stratification decisions such as who needs to be admitted to the hospital, and to which level of care intensity. The model was developed at Mass Gen Brigham (MGB), using Chest X-Ray (CXR) images, vital signs, demographic data, and lab values that were shown to be predictive of COVID-19 patient outcomes\textsuperscript{20–23}. CDS outputs a score, termed ‘CORISK’, that predicts oxygen support requirements, and that could be used as a decision support tool for triaging patients by front-line clinicians\textsuperscript{24–26}.

Healthcare providers have been known to prefer models that were validated on their own data\textsuperscript{27}. To date, most AI models have been trained and validated on ‘narrow’ data, often lacking in diversity\textsuperscript{28,29}. This results in a less
generalisable model. Even near-perfect, peer-reviewed, performance evaluation does not guarantee generalisability or a lack of over-fitting.

The study’s objective was to develop a robust model that could assist in triaging patients. We set to accomplish this by retraining the MGB CDS model with diverse data using an FL approach to develop a new model, EXAM (EMR Chest X-Ray Al Model), which would need to undergo further validation before clinical use. The hypothesis was that EXAM would perform better than local models, and it could generalise better across healthcare systems. Specifically, the model was trained to predict the aforementioned ‘CORISK score’ corresponding to a patient’s increasing oxygen requirements within two prediction windows, of 24 hours and of 72 hours following a patient’s initial presentation to the Emergency Department (ED).

EXAM is among the first FL COVID-19 models, and is the largest, and first multi-continent one. Chest X-ray was selected as the imaging input because it is widely available and commonly indicated by guidelines such as provided by ACR, Fleishnerr Society, the WHO, national thoracic societies, national health ministry COVID handbooks and radiology societies throughout the world. Accessing diverse data without the requirement of centralising data can also be enabled by methods such as Transfer Learning. FL was preferred by the authors due to its ability to rapidly launch centrally orchestrated experiments with improved traceability of data and assessment of algorithmic changes and impact.

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. FL has already shown promise in recent medical imaging applications including in COVID-19 analysis. A notable example is a mortality prediction model in patients infected with SARS-COV-2 that uses clinical features, albeit it limited in terms of number of modalities and scale.
Results

To analyze the stability of our results, we repeated three runs of local training and FL on different randomized data splits. 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). It also resulted in a 38% generalisability 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. 3). For the prediction results of 72h oxygen treatments (shown in Extended Data Fig. 7), the best global model training resulted in an average performance improvement of 18% compared to locally trained models, while generalisability of the global model improved on average by 34%.

The EXAM model was tested at three independent sites (Data Table 1). We present the ROC curves and confusion matrices for the largest data set, from CDH, in Fig. 5. The operating point was set to discriminate between non-mechanical ventilation (MV) treatment and MV treatment (or death). The FL trained model achieved an average AUC of 0.944 and 0.924 for 24/72h prediction tasks respectively, that exceeded the average performance among sites used in training EXAM. For prediction of MV in 72h, our model had a low false-negative rate of 7.1%.

When investigating a partial weight-sharing scheme\textsuperscript{49,50}, we showed that models can reach a comparable performance even when only 25% of the weight updates are shared (Fig. 4 and Methods section).

Discussion

To our knowledge, this study features the largest real-world healthcare FL experiment to date in terms of number of sites and number of data points used. The study encompassed 20 institutes, referred to as ‘client-sites’ and included over 16,000 cases (Extended Data Table 2). 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. EXAM 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 algorithm 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.

FL improved the model’s avg. AUC performance on all the different sites’ test sets, even for sites that had a large local training data set. For sites with small datasets, it would have been virtually impossible to build a performant deep learning model using only their local data. Furthermore, local models that were trained using unbalanced cohorts (e.g., mostly mild cases of COVID-19) markedly benefited from the FL approach (Extended Data Fig. 3 & 4). More importantly, the generalisability of the FL model was considerably increased over the locally trained model. We believe that is most likely due to a population or an age group that are under-represented in one hospital/region could be highly represented in another region (Extended Data Fig. 5 & 6). For example, children might be differentially affected by COVID-19, including disease manifestations in lung imaging51.

EXAM was designed in light of real-life clinical informatics circumstances, so a meticulous harmonization of the data input was not conducted, as seen in Fig. 1c/d and Extended Data Fig. 1. In-line with the CDS model, the features extracted from the medical record were carefully selected, to mitigate potential biases (Extended Data Table 1). Reported symptoms or clinical impressions were avoided. The CDS model was also trained on objective outcomes, which are practical to discern, being low-flow oxygen treatment, high-flow oxygen treatment, mechanical ventilation, and death (Extended Data Fig. 2). The application of high-flow oxygen treatment and ventilation require significant morbidity to be present, and it could thus transcend local practices52. We believe that these design considerations played a significant part in increasing the benefits from this FL approach and its impact on performance, generalisability and ultimately, the model’s usability. By participating in this study, the client-sites received access to the optimized AI model, EXAM (‘global FL
that can be further validated ahead of a potential future introduction into clinical care. The client-sites did not transfer data to a central repository but created a distributed data framework that can facilitate ongoing collaboration on AI model development and validation. This preservation of privacy encouraged participation of institutes who recognized the urgency to contribute during the COVID pandemic, but that would have been otherwise constrained by data governance considerations.

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’ or even the reconstruction of training images from the model gradients themselves. To counter these risks, security-enhancing measures were used, to mitigate risk in the event of data ‘interception’ during site-server communication. We experimented with techniques to avoid ‘interception’ of FL data, an added security feature that we believe could encourage more institutions to use FL. We thus validated previous findings (Fig. 4), showing that partial weight sharing, and other differential privacy techniques can successfully be applied in FL.

The validation results confirmed that the global model is robust. They strengthen the hypothesis that FL trained models are generalisable across healthcare systems. Thus, the EXAM study provides a compelling case for the use of predictive algorithms in covid-19 patient care, and the use of FL in model creation and testing. Further validation studies are needed before pursuing any regulatory approval of the model, and we are planning on validating EXAM prospectively in 'production' settings at MGB, as well as multiple other sites. However, at this time, the model is not approved by any regulatory agency and should only be used for research purposes.

A significant limitation of the work is the fact that despite the use of a large and diverse dataset in training the model, it has not been deployed and tested prospectively. We envision the future deployment of this model in the ED setting. It could be used 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 decision of whether to allocate patients to the medical ward or a higher level of care such as the ICU. In some cases, the risk score produced by
the model could be used as a reference point when deciding whether to release a patient back to be cared for in the community. We also envision using the model as a more sensitive population level metric, to help balance resources between regions, hospitals and departments.56.

Over 200 prediction models to support decision making in patients with COVID-19 have been published21. Unlike the majority of the articles published, focused on predicting mortality, we predict oxygen requirements that have implications for patient management. We also used patients with unknown SARS-COV-2 status, and so the algorithm could be used ahead of receiving an RT-PCR test result, making it more useful for a real-life clinical setting. The model’s image input, chest X-ray, is used in common clinical practice, in contrast with many models published that use chest Computed Tomography (CT) despite it being a non-consensual diagnostic modality. We also only used objective predictors, such as lab tests and vital signs, which is different from many of the published studies that leveraged subjective clinical impressions. The collection period and geographies used reflect varied incidence rates, and thus the ‘population momentum’ we encountered is more diverse. That could imply that the algorithm tested can be useful for populations with different incidence rates, and with different pre-test probabilities of suffering from COVID-19. In addition, bringing an algorithm to clinical practice is an expensive and difficult process, and thus having a federated project of this magnitude and breadth can provide a rapid, scalable way of bringing innovation into the clinic.

In the systemic sense, the main areas for development arising out of this collaboration are to streamline data access, local aggregation and preparation, in order to better leverage a network of sites participating in FL. Patient cohort identification and data harmonization are not novel issues in research and data science57, but can be further complicated, when using FL, given the lack of visibility on other sites’ data sets. There’s also a need for evolving our understanding of architectural considerations that will enable capturing more value out of FL, e.g., explicitly addressing the data domain shifts between the different participating sites in designing the model’s architecture.58. 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.42. Socio-economic status or ethnicity in an algorithm prototyped on a
homogenous population could enable algorithms to capture more diversity when using an FL approach, despite
being less useful when only leveraging more homogenous or and/or single-site data set. We would like to
compare the local and global model at the site where the local model was developed, and to understand where
the local model is best used versus the global model, and which known patient characteristics would encourage
the use of one or another (local or global). 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. A system of this nature was being set-up in MGB, at the time of study completion, and will be
used for prospective validation of the EXAM model. An additional aspect that we intend to investigate is the
potential for a ‘population drift’ due to different phases of disease progression, with associated radiographic and
clinical presentation. We believe that due to the diversity across 20 sites, this risk may be mitigated but we
would like to run serial assessment of the EXAM model in clinical practice and see the influence of data from
different cohorts on the model’s performance. As mentioned above, currently this model is not approved by any
regulatory body, and it should not be seen as such. Additional development that would foster these kinds of
large-scale collaborations, would be an improvement in the ability to predict each client-site’s contribution
towards improving the global FL model, which will help in client-site selection and prioritizing data acquisition
and annotation efforts in the future. The latter is especially important given the high costs and difficult logistics
of these large consortia endeavours, and the opportunity to capture diversity rather than the sheer quantity of
data samples.


16. (ESR), E. S. of R. What the radiologist should know about artificial intelligence – an ESR white paper. *Insights into Imaging* vol. 10 (2019).


### Data Table 1: Performance of the best global model, EXAM, on independent data sets.

**a.** Breakdown of patients by level of oxygen need at different time points across the 3 independent datasets, CDH, MVH and NCH. **b.** AUC of predicting level of oxygen need at 24 hr and 72 hrs at the 3 independent datasets (with 95% confidence. The AUC at NCH for ‘MV for 24 hrs could not be calculated as there were no mechanically ventilated patients.

#### a

<table>
<thead>
<tr>
<th>Site</th>
<th># Cases</th>
<th># Pos. Cases</th>
<th>Oxygen Device</th>
<th>RA</th>
<th>LFO</th>
<th>HFO_NV</th>
<th>MV &amp; DEATH</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDH</td>
<td>840</td>
<td>244</td>
<td>24 hours</td>
<td>608</td>
<td>162</td>
<td>48</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>72 hours</td>
<td>575</td>
<td>173</td>
<td>62</td>
<td>30</td>
</tr>
<tr>
<td>MVH</td>
<td>399</td>
<td>30</td>
<td>24 hours</td>
<td>356</td>
<td>36</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>72 hours</td>
<td>351</td>
<td>39</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>NCH</td>
<td>264</td>
<td>29</td>
<td>24 hours</td>
<td>237</td>
<td>23</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>72 hours</td>
<td>235</td>
<td>22</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

#### b

<table>
<thead>
<tr>
<th>Site</th>
<th>Prediction task</th>
<th>≥LFO</th>
<th>≥HFO_NV</th>
<th>≥MV</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDH</td>
<td>24 hours</td>
<td>0.925 (0.903, 0.945)</td>
<td>0.950 (0.926, 0.971)</td>
<td>0.956 (0.918, 0.984)</td>
</tr>
<tr>
<td></td>
<td>72 hours</td>
<td>0.902 (0.881, 0.924)</td>
<td>0.931 (0.905, 0.955)</td>
<td>0.938 (0.893, 0.927)</td>
</tr>
<tr>
<td>MVH</td>
<td>24 hours</td>
<td>0.904 (0.844, 0.954)</td>
<td>0.836 (0.620, 0.978)</td>
<td>0.964 (0.925, 1.000)</td>
</tr>
<tr>
<td></td>
<td>72 hours</td>
<td>0.887 (0.827, 0.940)</td>
<td>0.872 (0.663, 0.992)</td>
<td>0.988 (0.973, 0.997)</td>
</tr>
<tr>
<td>NCH</td>
<td>24 hours</td>
<td>0.895 (0.833, 0.950)</td>
<td>0.984 (0.957, 1.000)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>72 hours</td>
<td>0.904 (0.850, 0.949)</td>
<td>0.947 (0.890, 0.991)</td>
<td>0.931 (0.897, 0.959)</td>
</tr>
</tbody>
</table>
Fig. 1 | Data used in the EXAM Federated Learning study. a, EXAM included 20 different sites from around the globe. b, Number of cases that each institution or site contributed (client 1 being the largest site) c, CXR intensity distributions at each client site d, Age of patients included at each client-site showing the min. and max. ages (asterisks) and mean and standard deviation (length of bars).
Fig. 2 | Description of the EXAM Federated Learning study. a, Proposed model to predict a COVID risk score. 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 ≥ 0.25) 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 (≥ 0.75) 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.

c, Federated Learning using a client-server setup.
**Fig. 3 | Federated Learning vs. local training performance.** a, Test performance of models predicting 24h oxygen treatment trained on local data only (Local) versus the performance of the best global model available on the server (FL (gl. best)). b, Generalisability (average performance on other sites’ test data) as a function of a client’s dataset size (# cases). The average performance improved by 16% (from 0.795 to 0.920 or 12.5 percentage points) compared to locally trained models alone, while average generalisability of the global model improved by 38% (from 0.667 to 0.920 or 25.3 percentage points). Note, 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). Therefore, client 14 was also excluded from the computation of the baseline average generalisability numbers.
Fig. 4 | 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 learning. The Fig. shows 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.
Fig. 5 | Performance of the best global model on the largest independent data set. a, ROC Performance and confusion matrices on the independent dataset, CDH, predicting oxygen treatment at 24 hr. b, ROC Performance and confusion matrices on the independent dataset, CDH, predicting oxygen treatment at 72 hr. We show the ROCs for three different cut-off values $t$ of the CORISK score.
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 (e.g., the MGB ethics board, reference # 2020P002673). Since no patient data was transferred between any of the participants and the study was considered of no more than minimal risk to patients, the requirement of a full IRB process was largely waived according to the Ethical Principles and Guidelines for the Protection of Human Subjects of Research (the “Belmont Report”) and the requirements of the Health Insurance Portability and Accountability Act (HIPAA) of 1996.

Study Setting

The study included data from 20 institutions; 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\textsuperscript{59}.

**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. 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. Patient age and EMR feature distributions varied greatly between sites, as expected due to the differing demographics between globally distributed hospitals (Fig. 1d and extended Data Fig. 1).

After the initial training phase of the study was completed, the model was then validated on three independent data sets from Cooley Dickinson Hospital (CDH), Martha's Vineyard Hospital (MVH), and Nantucket Cottage Hospital (NCH), all of them in Massachusetts, USA. These hospitals had different patient population characteristics and were distinctly not part of the EXAM training sites. For this validation, patient data was included from March 2020 to February 2021, that satisfied the inclusion criteria (see Patient inclusion criteria below). The cohort size, labels and model inference results are summarized in Data Table 1.
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 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.

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 Extended Data Table 1. The variation of these features across different client-sites can be appreciated in Extended Data Fig. 1. The distribution of oxygen treatment using different devices at different client-sites is shown in Extended Data Fig. 2, which details the device usage at admission to the ED, and after 24-hour and 72-hour periods. 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 Extended Data Table 2. 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. 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\textsuperscript{61,62}. 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\textsuperscript{63}. The EXAM model was developed to help triage these patients.

**CORISK score in EXAM**

The EXAM model is based on a preliminary CDS, a patient outcome prediction model developed at MGB, that calculates a risk score termed “CORISK”\textsuperscript{30}.

EXAM was trained to predict this CORISK score\textsuperscript{30} corresponding to a patient’s oxygen needs within two prediction windows, 24 hours and 72 hours after initial presentation to the ED. We set the outcome labels of patients to 0, 0.25, 0.5, and 0.75 if the most intensive oxygen therapy the patient received in the prediction window was room air (RA), low-flow oxygen (LFO), high-flow oxygen (HFO)/non-invasive ventilation (NIV), or MV, respectively. 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. For EMR features, only the first values captured in the ED were used, and data pre-processing included de-identification, missing value imputation (using the MissForest algorithm\textsuperscript{64}), and normalization to zero-mean and unit variance. For CXR images, only the first one obtained in the ED was used. These CXR images were pre-processed to select the Anterior Position image and exclude lateral view images, and then scale to a resolution of 224x224. As shown in Fig. 2, the model fuses information from both the EMR features and CXR features (based on a modified ResNet-34 with spatial attention\textsuperscript{30,65} pre-trained on the CheXpert dataset)\textsuperscript{66}, and Deep & Cross network\textsuperscript{67}. 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. We used cross-entropy as the loss function and ‘Adam’ as the optimizer. The model was implemented in Tensorflow using the NVIDIA Clara Train SDK. 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 224x224.
Feature imputation & normalization

A MissForest algorithm 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. 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, or variations thereof. This algorithm can be realised 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.

A pseudo-algorithm of FL is shown in Extended Data Algorithm 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 $5\times10^{-5}$ and a stepwise learning rate decay with a factor 0.5 after every 40 epochs, which is important for the convergence of Federated Averaging. 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 when dealing with different clients in a non-independent and identically distributed (non-IID) setting, we found the best model performance to occur when keeping the pre-trained ResNet34 with spatial attention 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 $k(t)$ (Fig. 4), which was computed from all non-zero gradients $D_{W_k(t)}$ 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$^{72}$ that add random noise to the gradients or even to the raw data before feeding it to the network$^{73}$.

**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. 3 and Extended Data Fig. 8). 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 generalisability (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 24hr model and $r=0.62$, $p=0.003$, $df=16$ for the 72hr 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 (shown in Fig. 5), 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: $D = (AUC_1 - AUC_2)/s$, where $s$ is the standard deviation of the 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 Extended Data Table 3. 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\textsuperscript{73}.

**Observations**

**FL and dataset size**

We compared locally trained models with the global FL model on each client’s test data. For a client-site with a relatively small dataset, there are two typical ways to train a model: one is to train locally with its own data, the other is to apply a model trained on a larger dataset. It is shown in Extended Data Fig. 5 that these two ways are outperformed on all three tasks by the FL model significantly, indicating that the benefit for client-sites with small datasets is substantial.

Client-site (#16) provided an unbalanced dataset, with most patients experiencing mild disease severity, and with only a few severe cases. The improvement in prediction avg. AUC performance for the category with few cases was substantial; see Extended Data Fig. 3, $t \geq 0.5$ (categories $\geq$ high-flow oxygen device). The FL model achieved a higher *true positive rate* for the two positive (severe) cases at a markedly lower *false positive rate* compared to the local model, both shown in the receiver operating characteristic (ROC) plots and confusion matrices. The difference in dataset distribution for the two compared client-sites can be seen in Extended Data Fig. 4.

**Effect of COVID-19 status**

Extended Data Fig. 6 shows the performance of our model in predicting oxygen need at 24/72h for COVID positive/negative patients respectively. The global model performed well on both COVID positive and COVID negative patients. This is critical to the usability of the model, since PCR test results are usually not available at the time of ED disposition.
The study found the global model to be more robust compared to locally trained models when assessed across all client-sites' test data. Locally optimized models might provide improved performance on a client-site's own test data, but could result in a loss of generalisability. Local model selection heavily depends on the local validation set's quality and how well it represents the real test data's characteristics. In contrast, the global model, selected based on the averaged validation scores from each client-site, has better generalisability.

It is possible to train higher-performing models on a local dataset by tuning the training strategies more exhaustively, such as varying data augmentation, learning rate schedule, and data sampling methods. However, generalisability to other sites' data is expected to remain limited due to the lack of representative training data. Future approaches may incorporate automated hyperparameter searching, neural architecture search, and other automated machine learning (AutoML) approaches to find the optimal training parameters for each client-site more efficiently.

Slow or interrupted internet connectivity sometimes caused some clients' model updates to not be included in some rounds of the FL training. Such clients are commonly known as “stragglers”. While asynchronous FL algorithms might mitigate this issue, they usually require a special design of the algorithm or infrastructure to guarantee convergence, a topic that has been better studied and validated in synchronous settings.

Known issues of BN in FL motivated us to fix our base model for image feature extraction 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-IID.

Recent works on privacy attacks within the FL setting have raised concerns on data leakage during model training. Meanwhile, the protection algorithms are still under-explored and constrained by multiple factors. While differential privacy algorithms show good protection, they may weaken the model’s performance. The encryption algorithms, such as Homomorphic Encryption shall maintain the performance but may significantly increase the model 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$^{39,49,55}$.

As data was not centralized, it is not readily accessible and remains under the custody of the collaborating institutes. Given that, any future analysis of the results, beyond what was derived and collected, is limited.

A final, but important limitation to all machine learning models is that they are limited by the quality of the training data. Institutions interested in deploying these algorithms for clinical care need to understand the inherent biases in the training. For example, the ground truth data used in the training of the EXAM model was 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 oxygen need predictions made by this model. It is important to study the failure cases of AI models. We did not have sufficient information for the generation of significant statistics to test hypotheses regarding failure causes, but we postulate potential hypotheses that we can test in the future. It appears that for the high-performing sites, most but not all the failure cases fall into two categories: 1) bad quality of input data, e.g. too many missing data or severe motion artifact in CXR; 2) out-of-distribution data, e.g. a very young patient or a patient who's condition deteriorates very quickly. Further studies are needed to evaluate these hypotheses and study failure cases in more detail. While we could not do this in this study due to data access limitations, we did however study the failure cases from the largest independent test site, CDH. For low-performance sites, the performance was strongly correlated with the insufficient number of local samples, and thus the analysis of the failure cases is not so robust or meaningful. In Fig. 5, we showed the testing result at CDH, including the confusion matrix of the prediction of 24 hour MV, and in Extended Data Fig. 8, we present the two false-negative cases from CDH. One case had many missing EMR data inputs, the other one had a CXR with motion artifact and some missing EMR features.
**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.

**Code availability**

The model, code for training, validating testing the model, readme file, installation guideline, and license files can be accessed at NVIDIA NGC [39]:

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 (AIMTU) National Health Insurance Administration, Taiwan and Ministry of Science and Technology, Taiwan 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.

**Author information**

These authors contributed equally: Ittai Dayan, Holger Roth, Aoxiao Zhong, Fiona J Gilbert, Quanzheng Li, Mona G. Flores

**Affiliations**

1. **MGH Radiology and Harvard Medical School, Boston, MA, USA**
   - Keith Dreyer & Ittai Dayan

2. **NVIDIA, Santa Clara, CA, USA**
   - Holger Roth, Ahmed Harouni, Anas Abidin, Andrew Liu, CK Lee, Colleen Ruan, Daguang Xu, Eddie Huang, Griffin Lacey, Jesse Tetreault, iahui Guan, Kristopher Kersten, Nicola Rieke, Pedro Mario Cruz Silva, Mona G. Flores, Abood Quraini, Andrew Feng, Colin Compas, Deepeksa Bhatia, Isaac Yang, Mohammad Adil & Yuhong Wen
3. Center for Advanced Medical Computing and Analysis, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
   Aoxiao Zhong, Dufan Wu, Hui Ren, Xiang Li & Quanzheng Li

4. San Diego VA Health Care System, San Diego, CA, USA
   Amilcare Gentili

5. Department of Neurosurgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA
   Anthony Beardsworth Costa & Young Joon Kwon

6. Radiology & Imaging Sciences / Clinical Center, National Institutes of Health, Bethesda, MD, USA
   Bradford J. Wood

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

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.
   Chih-Hung Wang

9. Center for Research in Biological Systems, University of California, San Diego, CA, USA
   Chun-Nan Hsu

10. Diagnósticos da América SA (DASA), Brazil
    Felipe Campos Kitamura, Gustavo César de Antônio Corradi, Matheus Ribeiro Furtado de Mendonça & Vitor de Lima Lavor

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

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

13. Self-Defense Forces Central Hospital, Tokyo, Japan
    Hirofumi Obinata, Shuichi Kawano, Hisashi Sasaki, Hitoshi Mori & Tatsuya Kodama

14. Center for Intelligent Imaging, 2Department of Radiology and Biomedical Imaging, University of California, San Francisco, California, USA
    Jason C. Crane, Pablo F. Damasceno, Christopher P. Hess, Jae Ho Sohn & Sharmila Majumdar

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

16. Department of Radiology, NIHR Cambridge Biomedical Resource Centre, University of Cambridge, Cambridge, UK
    Josh D Kaggie, Fiona J Gilbert & Sarah Hickman

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
Keith Dreyer, Marcio Rockenbach, Varun Buch, Bernardo Bizzo and Evan Leibovitz

19. Sheikh Zayed Institute for Pediatric Surgical Innovation, Children's National Hospital, Washington, DC, USA
Carlos Tor Diez & Marius George Linguraru

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

21. Lunenfeld-Tanenbaum Research Institute, Toronto, Canada
Meena AbdelMaseeh

22. MeDA Lab and Institute of Applied Mathematical Sciences, National Taiwan University, Taipei, Taiwan
Pochuan Wang & Weichung Wang

23. Center for Interventional Oncology, National Institutes of Health, Bethesda, MD, USA
Sheng Xu & Sheridan Reed

24. Research Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand, Center for Artificial Intelligence in Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand
Sira Sriswasdi

25. Department of Internal Medicine, School of Medicine, Kyungpook National University, Daegu, South Korea
Soo Young Park, Won Young Tak & Yu Rim Lee

26. Departments of Radiology, Medical Physics, and Biomedical Engineering, The University of Wisconsin-Madison School of Medicine and Public Health, Madison, WI, USA
Thomas M. Grist

27. 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
Watsamon Jantarabenjakul & Thanyawee Puthanakit

28. Medical Review and Pharmaceutical Benefits Division, National Health Insurance Administration, Taipei, Taiwan
Weichung Wang & Chiu-Ling Lai

29. Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
Xihong Lin

30. Department of Radiology, NIHR Cambridge Biomedical Resource Centre, Cambridge University Hospital, Cambridge, UK
Andrew N Priest
31. Department of Radiology and Imaging Sciences, National Institutes of Health, Bethesda, MD, USA and National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
   Baris Turkbey

32. 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
   Benjamin Glicksberg

33. Department of Internal Medicine, Catholic University of Daegu School of Medicine, Daegu, South Korea
   Byung Seok Kim

34. Planning and Management Office, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan, R.O.C.
   Chia-Jung Hsu & Chia-Cheng Lee

35. 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.
   Chin Lin

36. Department of Neurosurgery, NYU Grossman School of Medicine, New York, NY, USA
   Eric K Oermann

37. MOST/NTU All Vista Healthcare Center, Center for Artificial Intelligence and Advanced Robotics, National Taiwan University, Taipei, Taiwan
   Li-Chen Fu

38. Division of General Internal Medicine and Geriatrics (Fralick), Sinai Health System, Toronto, Canada
   Mike Fralick

39. Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand
   Peerapon Vateekul

40. Schwartz/Reisman Emergency Medicine Institute, Sinai Health, Toronto, ON, Canada and Department of Family and Community Medicine, University of Toronto, Toronto, ON, Canada
   Shelley L. McLeod

41. Department of Medicine, NIHR Cambridge Biomedical Resource Centre, University of Cambridge, Cambridge, UK
   Stefan Graf

42. 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
   Stephanie Harmon
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 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, Deepeksa 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, Josh D Kagge, Jung Gil Park, Keith Dreyer, Marcio Aloisio Bezerra Cavalcanti Rockenbach, Marius George Linguraru, Masoom A. Haider, Meena AbdelMaseeh, Pablo F.
Corresponding authors

Correspondence to Mona G. Flores

Ethics declarations

Competing interests


I.D. is the CEO of Rhino HealthTech Inc., and owns stock in the company.

J.G. declared ownership of NVIDIA Stock.

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.

**Additional information**

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

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | 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. 2 | 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. 3 | Comparison of the Federated Learning trained vs. locally trained model at a site with unbalanced data and mostly mild cases. a, ROC Performance. b, Confusion matrices on the test data at site 16 predicting oxygen treatment at 72h using the locally trained model. c, 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 CORISK score.
Extended Data Fig. 4 | 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. 5| Effect of small data set size on ROC.** ROC of the locally trained model and the mean ROC of models trained on larger datasets in comparison to the best global model to predict oxygen treatment at 72h, using the test data at client-site 12 - a client-site with relatively small dataset. The Mean ROC is calculated based on models trained on 5 large datasets from client-sites in the Boston area, with the grey-area showing the standard deviation of the ROCs. We show the ROCs for three different cut-off values $t$ of the CORISK score.
Extended Data Fig. 6 Effect of large data set size on ROC. 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 CORISK score.
Extended Data Fig. 7 | 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)).

Generalisability (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 generalisability of the global model improved by 34% (from 0.669 to 0.899 or 23.0 percentage points).
Extended Data Fig. 8| Failures cases at an independent test site. EMR features and CXRs from two failure cases at CDH.
## Extended Data Table 1 | EMR (electronic medical record) data used in the EXAM study

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcategory</th>
<th>Component Name</th>
<th>Definition</th>
<th>Units</th>
<th>LOINC Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic</td>
<td>-</td>
<td>Patient Age</td>
<td>-</td>
<td>Years</td>
<td>30525-0</td>
</tr>
<tr>
<td>Imaging</td>
<td>Portable Chest X-Ray</td>
<td>-</td>
<td>AP or PA Portable Chest X-ray</td>
<td>-</td>
<td>36554-4</td>
</tr>
<tr>
<td>Lab Value</td>
<td>C-Reactive Protein</td>
<td>C Reactive Protein</td>
<td>Blood C-Reactive Protein Concentration</td>
<td>mg/L</td>
<td>1988-5</td>
</tr>
<tr>
<td>Lab Value</td>
<td>CBC (Complete Blood Count)</td>
<td>Neutrophils</td>
<td>Blood Absolute Neutrophils</td>
<td>10^9/L</td>
<td>751-8</td>
</tr>
<tr>
<td>Lab Value</td>
<td>D-Dimer</td>
<td>D-Dimer</td>
<td>Blood D-Dimer Concentration</td>
<td>ng/mL</td>
<td>7799-0</td>
</tr>
<tr>
<td>Lab Value</td>
<td>Lactate</td>
<td>Lactate</td>
<td>Blood Lactate Concentration</td>
<td>mmol/L</td>
<td>2524-7</td>
</tr>
<tr>
<td>Lab Value</td>
<td>LDH (Lactate Dehydrogenase)</td>
<td>LDH</td>
<td>Blood Lactate Dehydrogenase Concentration</td>
<td>U/L</td>
<td>2532-0</td>
</tr>
<tr>
<td>Lab Value</td>
<td>Metabolic Panel</td>
<td>Creatinine</td>
<td>Blood Creatinine Concentration</td>
<td>mg/dL</td>
<td>2160-0</td>
</tr>
<tr>
<td>Lab Value</td>
<td>Procalcitonin</td>
<td>Procalcitonin</td>
<td>Blood Procalcitonin Concentration</td>
<td>ng/mL</td>
<td>33959-8</td>
</tr>
<tr>
<td>Lab Value</td>
<td>Metabolic Panel</td>
<td>eGFR</td>
<td>Estimated Glomerular Filtration Rate</td>
<td>mL/min/1.73m2</td>
<td>69405-9</td>
</tr>
<tr>
<td>Lab Value</td>
<td>Troponin</td>
<td>Troponin-T</td>
<td>Blood Troponin Concentration</td>
<td>ng/ml</td>
<td>67151-1</td>
</tr>
<tr>
<td>Lab Value</td>
<td>Hepatic Panel</td>
<td>AST</td>
<td>Blood AST Concentration</td>
<td>IU/L</td>
<td>1920-8</td>
</tr>
<tr>
<td>Lab Value</td>
<td>Metabolic Panel</td>
<td>Glucose</td>
<td>Blood Glucose Concentration</td>
<td>mg/dL</td>
<td>2345-7</td>
</tr>
<tr>
<td>Vital Sign</td>
<td>-</td>
<td>Oxygen Saturation</td>
<td>Oxygen Saturation</td>
<td>%</td>
<td>59408-5</td>
</tr>
<tr>
<td>Vital Sign</td>
<td>-</td>
<td>Systolic Blood Pressure</td>
<td>Systolic Blood Pressure</td>
<td>mmHg</td>
<td>8480-6</td>
</tr>
<tr>
<td>Vital Sign</td>
<td>-</td>
<td>Diastolic Blood Pressure</td>
<td>Diastolic Blood Pressure</td>
<td>mmHg</td>
<td>8462-4</td>
</tr>
<tr>
<td>Vital Sign</td>
<td>-</td>
<td>Respiratory Rate</td>
<td>Respiratory Rate</td>
<td>breaths per minute</td>
<td>9279-1</td>
</tr>
<tr>
<td>Vital Sign</td>
<td>-</td>
<td>COVID PCR test</td>
<td>PCR for RNA [not used as input to model]</td>
<td>-</td>
<td>95425-5</td>
</tr>
<tr>
<td>Vital Sign</td>
<td>Oxygen Device used at Emergency Department (ED)</td>
<td>Ventilation, High-flow/NIV, Low-flow, Room Air</td>
<td>-</td>
<td>41925-9</td>
<td></td>
</tr>
<tr>
<td>Outcome</td>
<td>24Hr Oxygen Device</td>
<td>Ventilation, High-flow/NIV, Low-flow, Room Air</td>
<td>-</td>
<td>41925-9</td>
<td></td>
</tr>
<tr>
<td>Outcome</td>
<td>72Hr Oxygen Device</td>
<td>Ventilation, High-flow/NIV, Low-flow, Room Air</td>
<td>-</td>
<td>41925-9</td>
<td></td>
</tr>
<tr>
<td>Outcome</td>
<td>Death</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Outcome</td>
<td>Time of Death</td>
<td>-</td>
<td>-</td>
<td>Hours</td>
<td>-</td>
</tr>
<tr>
<td>Site</td>
<td># Cases</td>
<td># Pos. Cases</td>
<td># Neg. Cases</td>
<td>% Pos. Cases</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2994</td>
<td>1057</td>
<td>1937</td>
<td>35.3%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2825</td>
<td>139</td>
<td>2686</td>
<td>4.9%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2697</td>
<td>258</td>
<td>2439</td>
<td>9.6%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1786</td>
<td>618</td>
<td>1168</td>
<td>34.6%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1065</td>
<td>347</td>
<td>718</td>
<td>32.6%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>853</td>
<td>427</td>
<td>426</td>
<td>50.1%</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>724</td>
<td>168</td>
<td>556</td>
<td>23.2%</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>637</td>
<td>232</td>
<td>405</td>
<td>36.4%</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>565</td>
<td>342</td>
<td>223</td>
<td>60.5%</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>485</td>
<td>304</td>
<td>181</td>
<td>62.7%</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>400</td>
<td>400</td>
<td>0</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>346</td>
<td>346</td>
<td>0</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>213</td>
<td>114</td>
<td>99</td>
<td>53.5%</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>176</td>
<td>72</td>
<td>104</td>
<td>40.9%</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>102</td>
<td>102</td>
<td>0</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>99</td>
<td>99</td>
<td>0</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>74</td>
<td>49</td>
<td>25</td>
<td>66.2%</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>55</td>
<td>55</td>
<td>0</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>28</td>
<td>28</td>
<td>0</td>
<td>100.0%</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>24</td>
<td>15</td>
<td>9</td>
<td>62.5%</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>16148</strong></td>
<td><strong>5172</strong></td>
<td><strong>10976</strong></td>
<td><strong>32.0%</strong></td>
<td></td>
</tr>
</tbody>
</table>
Extended Data Table 3 | p-values and 95% CI of ROC comparisons of local training to Federated Learning shown in Extended Data Fig. 5. These ROC comparisons are for three different cut-off values \( t \) of the CORISK score

<table>
<thead>
<tr>
<th>p-values</th>
<th>( t \geq 0.25 )</th>
<th>( t \geq 0.5 )</th>
<th>( t \geq 0.75 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(95% CI improvement in AUC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Site 1</strong></td>
<td>0.02674 (0.0103, 0.0134)</td>
<td>0.03455 (0.0096, 0.0222)</td>
<td>0.04015 (0.0457, 0.0122)</td>
</tr>
<tr>
<td><strong>Site 4</strong></td>
<td>4.866e-05 (0.0277, 0.0358)</td>
<td>1.08e-05 (0.0465, 0.0850)</td>
<td>3.92e-05 (0.0614, 0.1264)</td>
</tr>
<tr>
<td><strong>Site 5</strong></td>
<td>3.005e-06 (0.0329, 0.0403)</td>
<td>5.098e-05 (0.04428, 0.0786)</td>
<td>0.0001362 (0.0473, 0.0928)</td>
</tr>
<tr>
<td><strong>Site 6</strong></td>
<td>1.717e-14 (0.0782, 0.0938)</td>
<td>5.816e-08 (0.1044, 0.1678)</td>
<td>2.357e-05 (0.0953, 0.1882)</td>
</tr>
<tr>
<td><strong>Site 8</strong></td>
<td>2.872e-10 (0.0520, 0.0618)</td>
<td>3.19e-06 (0.0570, 0.0906)</td>
<td>5.872e-05 (0.0575, 0.1095)</td>
</tr>
<tr>
<td><strong>Site 12</strong></td>
<td>0.08332 (0.0401, 0.0565)</td>
<td>0.0507 (0.006, 0.891)</td>
<td>0.02664 (0.0235, 0.1633)</td>
</tr>
</tbody>
</table>
Extended Data Algorithm 1 | Client-server based Federated Learning using the Federated Averaging algorithm \(^{53,58}\) as implemented in NVIDIA Clara Train SDK\(^{52}\)

Require: Number federated rounds \(T\).
Require: Number of local training iterations \(n_k\) for client \(k\).

1: procedure FEDERATED AVERAGING
2: Initialize model weights.
3: for Round \(t\) of \(T\) do
4: for client \(k\) of \(K\) do
5: Send current global model to client.
6: Evaluate global model on local validation data.
7: Initialize optimizer and momentums.
8: Perform training on local data.
9: Apply privacy-preserving scheme.
10: Send weight updates and validation scores to server.
11: Select locally best model based on validation score.
12: end for
13: Select best global model based on local validation scores.
14: Aggregate the client weight updates.
15: Update the global model weighted by local iterations \(n_k\).
16: end for
17: Exchanged models and evaluate on each client.
18: return Validation results.
19: return Final global model.
20: return Best global model.
21: return Locally best model for each client.