



Developing and Validating Multi-Modal Models for Mortality Prediction in COVID-19 Patients: a Multi-center Retrospective Study

Joy Tzung-yu Wu¹ · Miguel Ángel Armengol de la Hoz^{2,3,4} · Po-Chih Kuo^{2,5} · Joseph Alexander Paguio^{6,9} · Jasper Seth Yao^{6,9} · Edward Christopher Dee⁷ · Wesley Yeung^{2,8} · Jerry Jurado⁹ · Achintya Moulick⁹ · Carmelo Milazzo⁹ · Paloma Peinado¹⁰ · Paula Villares¹⁰ · Antonio Cubillo¹⁰ · José Felipe Varona¹⁰ · Hyung-Chul Lee¹¹ · Alberto Estirado¹⁰ · José Maria Castellano^{10,12} · Leo Anthony Celi^{2,13,14}

Received: 4 November 2021 / Revised: 15 May 2022 / Accepted: 8 June 2022
© The Author(s) 2022

Abstract

The unprecedented global crisis brought about by the COVID-19 pandemic has sparked numerous efforts to create predictive models for the detection and prognostication of SARS-CoV-2 infections with the goal of helping health systems allocate resources. Machine learning models, in particular, hold promise for their ability to leverage patient clinical information and medical images for prediction. However, most of the published COVID-19 prediction models thus far have little clinical utility due to methodological flaws and lack of appropriate validation. In this paper, we describe our methodology to develop and validate multi-modal models for COVID-19 mortality prediction using multi-center patient data. The models for COVID-19 mortality prediction were developed using retrospective data from Madrid, Spain ($N=2547$) and were externally validated in patient cohorts from a community hospital in New Jersey, USA ($N=242$) and an academic center in Seoul, Republic of Korea ($N=336$). The models we developed performed differently across various clinical settings, underscoring the need for a guided strategy when employing machine learning for clinical decision-making. We demonstrated that using features from both the structured electronic health records and chest X-ray imaging data resulted in better 30-day mortality prediction performance across all three datasets (areas under the receiver operating characteristic curves: 0.85 (95% confidence interval: 0.83–0.87), 0.76 (0.70–0.82), and 0.95 (0.92–0.98)). We discuss the rationale for the decisions made at every step in developing the models and have made our code available to the research community. We employed the best machine learning practices for clinical model development. Our goal is to create a toolkit that would assist investigators and organizations in building multi-modal models for prediction, classification, and/or optimization.

Keywords Multi-modal · Mortality prediction · COVID-19 · Multi-center

Introduction

Beginning as an outbreak of an unknown viral pneumonia in Wuhan, China, the coronavirus disease 2019 (COVID-19) pandemic has sparked numerous efforts to create predictive models. In particular, machine learning methods hold great promise because they provide the opportunity to

combine and use features from multiple modalities available in electronic health records (EHR), such as imaging and structured clinical data, for downstream prediction tasks. At present, there are hundreds of papers in preprint servers and medical journals employing machine learning methodologies in an attempt to bridge the gaps in the diagnosis, triage, and management of COVID-19; eight of them have integrated both radiological and clinical data [1–8].

However, most of these studies were found to have little clinical utility, producing a credibility crisis in the realm of artificial intelligence in healthcare. A recent review by Roberts et al. found that, after screening more than 400 machine learning models using various risk and bias assessment tools, none of the evaluated machine learning models had sufficiently fulfilled all of the following: (1)

Joy Tzung-yu Wu, Miguel Ángel Armengol de la Hoz and Po-Chih Kuo contributed equally on this work.

Leo Anthony Celi and José Maria Castellano are co-senior authors.

✉ Po-Chih Kuo
kuopc@cs.nthu.edu.tw

Extended author information available on the last page of the article

documentation of reproducible methods, (2) adherence to best practices in the development of a model, and (3) external validation that could justify claims of applicability [9].

Furthermore, the question remains as to how useful these predictive models actually are to other institutions to which these models were not customized [10]. While machine learning models offer the potential for a more accurate prediction of clinical outcomes within a specific context, these models were usually trained using data from a single institution and are unable to identify differences in contexts when employed in other settings [11]. This problem raises the need for validation not just in the neighboring center, but in other types of centers, states, or even countries, where patient demographics, standards of care, institutional policies may largely differ. In addition, these models need to be constantly updated because the contexts in which these models were trained and approved for use may be significantly different when used at present day [12–15]. Finally, beyond concerns about the reproducibility and generalizability of machine learning models is the issue of the lack of explainability, in which models may draw spurious associations between confounding imaging features and the outcome of interest [12, 16]. DeGrave et al. attempted to assess the trustworthiness of recently published machine learning models for COVID-19 by using explainable AI technology to determine which regions of chest X-rays (CXR) these models used to predict outcomes [12]. Surprisingly, they found that in addition to highlighting lung regions, the evaluated models used laterality marks, CXR text markers, and other features that provide no pathologic basis for distinguishing between COVID-positive and COVID-negative studies [12]. In other words, it was discovered that these models used shortcuts, further underscoring concerns about their applicability.

Therefore, we believe that some of the ways investigators can address the questions surrounding the credibility of a machine learning model are (1) to state the clinical context, which include patient demographics, geography, and timeframe, of the training and testing datasets that were used (2) to provide the resources for other centers to create or fine tune models specific to their contexts, (3) to be explicit about the appropriate level of the model's generalizability based on results of external validation studies, (4) to explore strategies that either build in and/or evaluate the explainability of models, and (5) to externally validate the performance of the model on different subpopulations of the sample and explore the fairness of the model in under-represented patient groups.

In our case, we present our efforts to develop three machine learning models for predicting 30-day mortality among hospitalized patients with COVID-19: (1) a structured EHR-based model, (2) a CXR-based model, and (3) an

EHR-CXR fusion model. All three models were developed using a multi-center dataset from Madrid, Spain. We aim to investigate how the performance of each of these models differed when validated on two external unseen single-center datasets from different countries (the USA and Republic of Korea). In addition, we will flag and detail why certain modeling design decisions were made, including the difficulties and trade-offs of these decision-making processes. The Checklist for Artificial Intelligence in Medical Imaging [17] is used to report our study designs and findings. We have made the code and other resources to reproduce our model training process available to the research community. This work has the potential to inform triage allocation when demand exceeds hospital capacity or may aid in the prediction of the level of care, guiding inpatient assignment of patients. We hope our work would serve as a toolkit that future investigators could use, adapt, and retrain models using data from their own institutions. Ultimately, our goal is to provide other institutions the opportunity to leverage machine learning technology to predict the mortality of their patients with COVID-19 and customize these models to meet their individual institution's needs.

Materials and Methods

Study Objectives

We used retrospective data from Hospitales de Madrid to build three machine learning models that taking input from (1) only structured EHR data, (2) the first CXR image, and (3) both the EHR data and CXR image for predicting COVID-19 patient's mortality at 30 days from hospital admission, as illustrated in Fig. 1. Our aims are (1) to investigate if modeling with features from both EHR and CXR image data result in improved mortality prediction performance and (2) to investigate if modeling with features from both EHR and CXR image data result in more consistent model performance on external test sets (from Hoboken, New Jersey, USA and Seoul, Republic of Korea).

Datasets

Three datasets of patients with confirmed SARS-CoV-2 infection from three different countries were used in our study (see details under “Case Definitions”). The hospital mortality outcomes came from the source EHRs [18, 19]. Table 1 documents the number of patients included and/or excluded for different reasons. We used the dataset from Hospitales de Madrid (HM), a network of 17 hospitals in Madrid, Spain, for all model training and hyperparameter tuning experiments. The dataset is accessible via credentialed and HIPAA-compliant approvals from <https://www>.

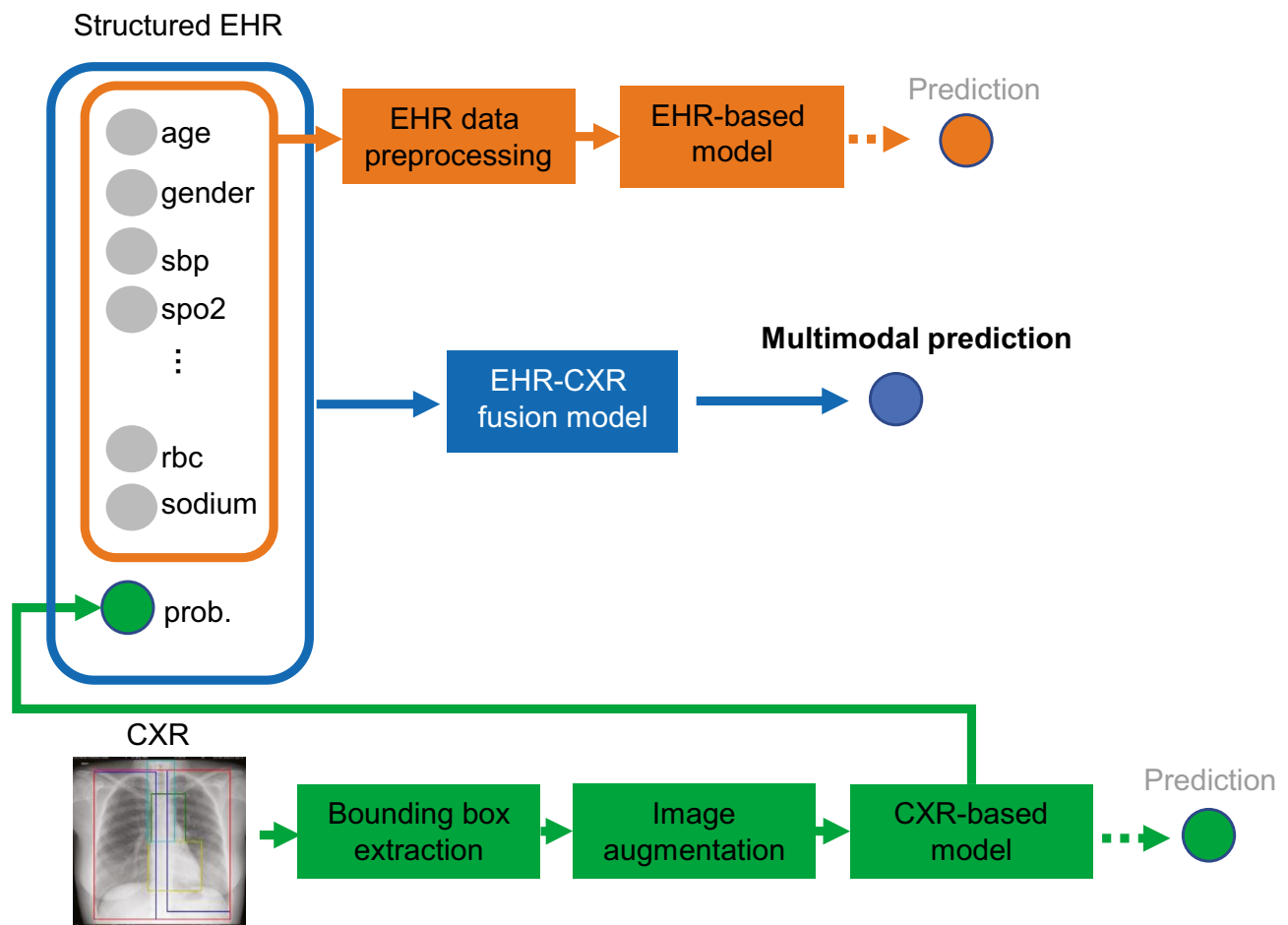


Fig. 1 The proposed multi-modal models for mortality prediction. The extracted EHR data were first preprocessed and then used to train the EHR-based model. For the CXR-based model, an anatomical bounding box extraction pipeline was used to automatically extract the coordinates for the left lung, right lung, mediastinum, and trachea anatomies from each of the CXR images. The CXR images with aug-

[hmhospitals.com/coronavirus/covid-data-save-lives/english-version](https://hospitals.com/coronavirus/covid-data-save-lives/english-version). External datasets from Hoboken University Medical Center (HUMC), USA and Seoul National University

mentation were then used to train the CXR-based model. The probability computed from the CXR-based model along with EHR data were used to train the proposed EHR-CXr fusion model, by which the final prediction was generated. The predictions from the EHR- and CXR-based models were also generated for the comparison

Hospital (SNUH), Republic of Korea were used for validating and evaluating the trained models. Table 2 describes the basic clinical characteristics of patients in each of the three

Table 1 High level descriptive summary of datasets used in this study

Dataset name	Data split	Inclusion criteria	Exclusion criteria	Size (number included/all)*
Madrid	fourfold training and internal validation for building and tuning the models	Multi-centered hospital network, Madrid, Spain, from 12/2019 to 06/2020 [18]	Under age (< 16)	$N=14$
			Missing admission time	$N=85$
			Missing admission chest X-ray	$N=820$
Hoboken	Test (external validation)	Community hospital, Hoboken, NJ, USA, from 03/2020 to 04/2020 [19]	Under age (< 16)	$N=0$
			Missing admission time	$N=0$
			Missing admission chest X-ray	$N=41$
Seoul	Test (external validation)	Academic tertiary hospital, Seoul, Republic of Korea, from 1/1/2020 to 12/31/2020	Under age (< 16)	$N=16$
			Missing admission time	$N=0$
			Missing admission chest X-ray	$N=5$

*These are unique patients

Table 2 Summary of clinical characteristics for the 3 different datasets used in the study

Characteristics	Madrid			Hoboken			Seoul		
	Alive	Expired	<i>p</i> value	Alive	Expired	<i>p</i> value	Alive	Expired	<i>p</i> value
<i>n</i>	1439	189		114	87		310	5	
Age (mean)	65.7	79.6	<0.001	61.9	69.1	0.003	45.7	64.0	0.053
Female (%)	41.3	28.6	<0.001	48.2	32.2	0.032	48.4	0	0.062
30-day mortality (%)	88.4	11.6	<0.001	56.7	43.3	<0.001	98.4	1.6	<0.001
Diabetes (%)	16.1	24.1	0.008	39.4	35.6	0.682	11.0	40.0	0.103
Hypertension (%)	6.1	10.2	0.055	54.3	55.2	0.974	13.9	80.0	0.002
Hyperlipidemia (%)	26.0	35.3	0.01	32.5	34.5	0.880	6.1	20.0	0.283
Congestive heart failure (%)	4.3	7.0	0.156	16.7	14.9	0.891	1.6	20.0	0.093
Ischemic heart disease (%)	6.0	13.4	<0.001	16.7	14.9	0.891	2.3	20.0	0.122
Stroke (%)	2.6	7.5	<0.001	2.6	4.6	0.469	2.6	0.0	-
COPD (%)	4.7	8.0	0.083	9.6	10.3	0.941	0.6	20.0	0.047
CKD (%)	5.1	11.8	<0.001	7.0	20.7	0.008	1.3	20.0	0.078
Chronic liver disease (%)	0.6	4.8	<0.001	0.9	0.0	1.000	1.0	0.0	-
Active cancer (%)	4.0	12.3	<0.001	6.1	3.4	0.519	0.0	0.0	-

COPD chronic obstructive pulmonary disease, *CKD* chronic kidney disease

different datasets. For more detailed patient clinical characteristics, please see Supplementary Table 1.

Case Definitions

Madrid From all patients with COVID-19 (N = 2547) admitted at Madrid, the vast majority of patients have been diagnosed by positive PCR. However, during the months of March–April 2020, when there was no PCR test, the diagnosis was made by clinical and/or radiological signs from an CXR and symptoms compatible with bilateral pneumonia.

Hoboken Data of all patients with COVID-19 (N = 242) admitted at the Hoboken University Medical Center until April 11, 2020, were retrospectively collected on April 21, 2020. COVID-19 was confirmed in all patients using quantitative real-time reverse transcription polymerase chain reaction for SARS-CoV-2 RNA. Data for patients who did not meet the primary outcome were excluded on the 30th day of admission.

Seoul Data of all patients with COVID-19 (N = 336) admitted at Seoul are patients diagnosed with COVID-19 (PCR confirmed) from Seoul clinical data warehouse (CDW) and admitted to the intensive care unit at Seoul.

Ethical Statements

Separate IRBs and data use agreements were independently obtained from the data controller and the ethics board of the source institutions for the three different datasets to conduct

this study. Data access to different datasets by researchers in this study is given via an as needed basis after the researchers and their institutions signed the relevant data use agreement. The purpose of the study is non-commercial and the ethical statements for the following datasets are the following:

Madrid CEIm Ref No. 20.05.1627-GHM Title of the Protocol: Clinical course and outcomes of severe and critical COVID-19 patients on interleukin-6 inhibitors: a retrospective cohort study; protocol identification: Covid-IL6; IRB Sponsor: Fundación de Investigación HM Hospitales.

Hoboken This patient population was previously reported by Yao et al. and the study protocol was approved and was granted a waiver of informed consent by the hospital board on April 15, 2020 [19]. Data extraction, collection, and analyses and external model validation on this dataset were performed by two trained physicians from HUMC (JAP and JSY).

Seoul IRB No. H-2007–065-1140 from Seoul National University Hospital, Republic of Korea.

Data Extraction

We included both comorbidities and lab EHR variables for the EHR-based and the EHR-CXR fusion models. Variables (categorical) included for comorbidities are diabetes, hyperlipidemia (HLD), hypertension (HTN), ischemic heart disease (IHD), chronic kidney disease, chronic obstructive pulmonary disease (COPD), asthma, cancer, chronic liver disease, stroke, congestive heart failure (CHF), and

dementia. Lab variables (numeric) include lactate dehydrogenase (LDH), hemoglobin, mean corpuscular volume (MCV), neutrophil percentage, mean neutrophil, lymphocyte percentage, mean lymphocyte, mean leukocyte, mean platelet volume, mean platelet, C-reactive protein (CRP), mean corpuscular hemoglobin (MCH), aspartate aminotransferase (AST), alanine aminotransferase (ALT), activated partial thromboplastin time (APTT), D-dimer, prothrombin activity, international normalized ratio (INR), glucose, sodium, potassium, blood urea nitrogen (BUN), and creatinine. Images were included and excluded as per Table 1.

Structured EHR Data Preprocessing

Deidentification The following identifiers of the individual or of relatives, employers, or household members of the individual were removed for de-identifying the dataset: names; all geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent codes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the US, all elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages elements may be aggregated into a single category of age 90 or older; telephone numbers; vehicle identifiers and serial numbers, including license plate numbers; fax numbers; device identifiers and serial numbers; email addresses; web Universal Resource Locators (URLs); social security numbers; internet Protocol (IP) addresses; medical record numbers; health plan beneficiary numbers; account numbers; any other unique identifying number, characteristic, or code.

Outliers We cleaned up systolic blood pressure (systolic BP), heart rate, SPO2 and temperature to be within valid ranges (code available). The valid ranges used for temperature, SPO2, heart rate, and systolic BP are 30–45 C, 1–100%, 20–300 bmp, 20–240 mmHg, respectively.

Missing Values We removed labs that had more than 50% missing values in the model development dataset (Madrid). The remaining missing values are imputed following common data science procedures, where missing categorical values were filled by most frequent value imputation and continuous values by median imputation. Variables not available in Hoboken or Seoul but available in Madrid were filled with 0 s.

Data Normalization The categorical variables were transformed by a one-hot encoding method. The numerical variables were scaled based on percentiles across the whole

Madrid dataset with the robust scaler method in python package (scikit learn 0.21).

Training and Model Picking

A 121-layer Densely Connected Convolutional Network (DenseNet-121) [20] was used as the model architecture. Four different types of machine learning were tried in a tuning setting to select for the best EHR-based model. The CXR-based model building including 5 steps: (1) online (real-time) image augmentation during training; (2) online CXR feature extraction; (3) mortality classification layers; (4) optimization settings; (5) hyperparameter tuning and model selection, as described in supplementary. For the fusion model, we took a late fusion approach that uses the output probability from the CXR model as a feature along with the EHR features for the 30-day mortality classification.

Statistical Analysis

In all three datasets, we reported the point estimates and 95% confidence interval (95% CI) of the reported validation metrics: areas under the receiver operating characteristic curves (AUROC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, and accuracy. 95% CIs were computed by bootstrapping the scores of the predictions 1000 times. The optimal cut-off value was determined when the absolute value of the difference between the sensitivity and specificity values is minimum.

Model Evaluation

Internal validation was performed by averaging the results across the 4 folds using the best hyperparameters for the final models using the Madrid dataset. External validation using the Hoboken and Seoul datasets were respectively performed by two in-house Hoboken physicians and a data scientist, all of whom were uninvolved in model tuning.

Specifically, validation of the Hoboken dataset was performed by clinicians of the community hospital who had the appropriate credentials to access patient health information. To accomplish this task, the necessary code was developed by an external team of data scientists. This code is made available (Supplementary Table 6) to allow future researchers to replicate our methodology that allows inter-institutional collaboration while complying with data governance standards and protecting sensitive patient information.

Evaluating Model Fairness

A more complete analysis of biases in our models is not within the scope of our study, particularly since the datasets came from countries with completely different ethnic makeup.

However, as a baseline, performance results for male and female are reported separately on all three datasets to explore how the models' performance differ between the gender strata.

Evaluating Model Explainability

We used SHapley Additive exPlanations (SHAP) [21] to show the feature importance in our EHR-based and fusion model [21]. The SHAP method estimates differences between models with different feature subsets and calculates SHAP values representing the importance of each feature to overall model predictions. The features with larger absolute SHAP values are supposed to contribute more to the prediction. A more positive SHAP value for a feature corresponds to a higher model predicted likelihood. For the fusion model, the SHAP analysis helps to show

whether the CXR model's prediction is important for the final mortality prediction. For evaluating the explainability of the CXR model's prediction, we visualize where on the image the model attended to most by using Grad-CAM [22]. Grad-CAM computes the gradient of the prediction scores of the features generated by convolutional layers to reveal which locations in the image are most important.

Design Decisions and Reasons

To address the questions surrounding the credibility of a machine learning model, we documented design decisions and reasons for our study objectives, datasets, data pre-processing, training and model picking, model evaluation, and code packaging for testing.

Study objectives

Design decisions	Optimize for F1-score for all three models for the 30-day mortality prediction task and report all metrics including areas under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, and accuracy
Reasons	30-day mortality is chosen as the target outcome in accordance with clinical precedence [23–25]. The F1-score finds an equal balance between PPV (precision) and sensitivity (recall), which gives a better indication of model performance for unbalanced dataset (mortality is relatively rare compared to survival). Reporting all metrics allows assessment of how the models might perform in populations with a different COVID-19-related mortality distribution. With COVID mortality rates varying with time, model drift can be a real concern under different care delivery parameters during surges [26]

Datasets

Design decisions	We included only cases with admission CXRs and results of laboratory tests taken within the first 24 h of hospital admission in Tables 1 and 2. Cases were also excluded for missing admission time. Patients aged 16 or under are excluded and only frontal (AP or PA) images are included for the CXR-based and EHR-CXR fusion models
Reasons	Our clinical goal is to develop an early assessment algorithm. Admission time is needed to establish the 30-day mortality cut off for this study. Patients under 16 need more privacy protection (very rare and more easily re-identifiable) and their CXR imaging appearance (anatomically) and disease outcome distributions are very different. Not all CXR exam orders include lateral images hence they are not included as an input for the models

Image preprocessing

Design decisions	An anatomical bounding box (Bbox) extraction pipeline was used to automatically extract the coordinates for the left lung, right lung, mediastinum, and trachea anatomies from each of the frontal CXR images [27]. The extracted bounding boxes are reviewed and manually corrected as needed by clinicians (JAP, JSY, ECD). We used these anatomical Bboxes to create 4 additional versions for each image for augmentation, where in version (1) trachea Bbox was masked out with 0's, (2) trachea Bbox was replaced with random noise, (3) background and trachea Bboxes were masked out with 0's, and (4) background and trachea boxes were replaced with random noise. The original and augmented images are pre-saved as JPEGs without resizing at this stage. During training, when the hyperparameter for "augment_bbox" is set to true, a random version of each image (including possibly the non-augmented version) is drawn to teach the model in each epoch (see Fig. 2)
Reasons	As compared to simply post hoc assessing the explainability of models with Gradient-weighted Class Activation Mapping (Grad-CAM), we tried to force the CXR model to learn features from key CXR anatomies that should be relied on more heavily for prediction during the model training stage as well [27]. Non-augmented image examples are also used in the training so that the model can handle non-augmented CXR images too at inference (i.e., clinical deployment setting). Doing the Bbox augmentation offline not only makes training faster but also more deterministic. We left input size for images as a tuning parameter that is dependent on the pre-trained model teacher

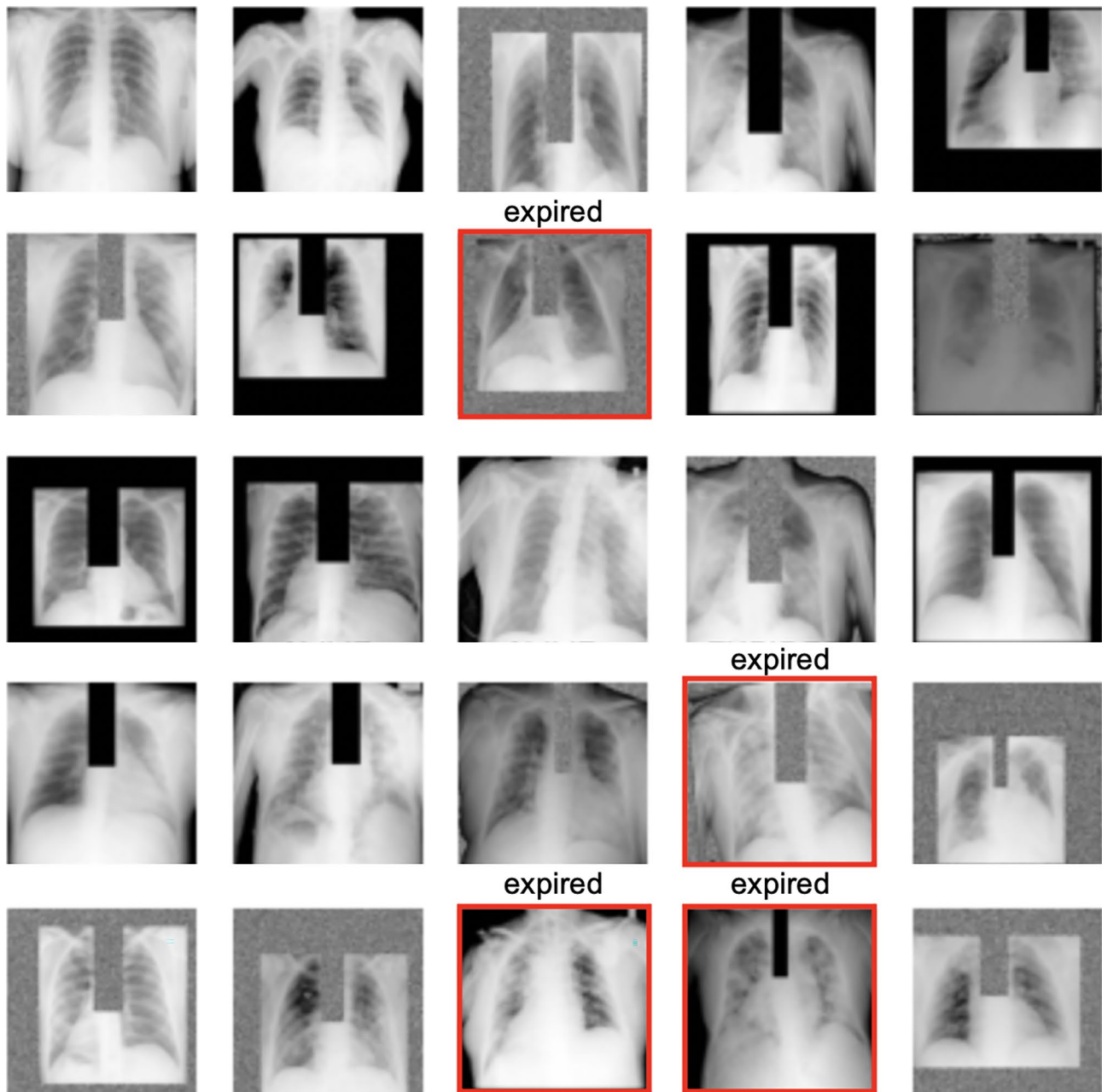


Fig. 2 A random sample of images shown to teach the model where at least 1–2 positive mortality (expired) cases are shown to the model in each batch

Training and model picking

All models

Design decisions

The whole of Madrid dataset was randomly divided into four subsets in order to conduct a fourfold cross-validation training strategy to select for the best model (by F1-score) for each of the three model types. We ensured similar numbers of mortality cases in each split and the same four-way split was used for all experiments. Finally, we trained each of the 3 models on all of Madrid data once we identified the best hyperparameters from the fourfold cross-validation hyperparameter tuning experiments. The final models are then validated on the two external datasets (Hoboken and Seoul). See supplementary materials for the details of model training and picking for EHR, CXR, and fusion models. All models are trained on Google Cloud TPUs via Colab notebooks. Code for both training with paid and free TPUs are available. Software packages used were tensorflow == 2.4.1, sklearn-pandas == 1.8.0, xgboost == 0.90. To ensure repeatability, a random seed of 2020 was used for all experiments

Training and model picking

All models

Reasons In this setup, for each experiment, 3 folds are combined and used for training and the fourth fold is used for validation, whereby each individual data subset gets equal opportunities to validate models. Model parameters that perform best on one validation subset might just be “lucky.” Rotating the validation subset and picking model parameters that perform the best on average across all subsets helps in selecting a model that has hopefully learned more reliable features and may generalize better on external validation sets. Similar number of positive mortality cases (expired patients) in each split makes the validation set more likely to be equally difficult. We had to use the whole Madrid dataset for model development (training and validation); otherwise, the number of positive cases (mortality) would be too small for tuning. We used only open sourced python packages so that others can easily re-use and build on our work with no cost barriers

EHR-based model

Design decisions Four different types of machine learning algorithms (logistic regression, random forest, gradient boosting, and XGBoost) implemented in scikit learn 0.21 were tried in a tuning setting to select for the best EHR-based model. A randomized grid-search method was used to sample different hyperparameter settings from prespecified ranges for each optimization experiment, as shown in Supplementary Table 3

Reasons The goal of this modeling is not causality analysis but simply to select a model that performs the best for the given dataset and prediction task. We picked the four most common machine learning algorithms suitable for modeling tabular data and tuned their hyperparameters

CXR-based model**Step 1: Online (real-time) image augmentation during training**

Design decisions CXR images were randomly flipped vertically (left–right) and brightness adjusted (0–0.05). Together with the preprocessed anatomical Bbox augmentation, a random set of CXRs used for training the model is illustrated in Fig. 2. Both the online and offline augmentations are only used during training and not during internal and external validation of models

Reasons The goal of image augmentation is to automatically increase training sample variety so that the model can learn to discern features that are more generalizable for the downstream prediction task. This step is particularly important if the training dataset is small. The online augmentations (flip and brightness) try to simulate how variations under which CXRs can be taken in real life might alter the image appearance. Only small augmentation ranges are chosen so that the CXR images remain radiologically interpretable. Augmentation is not used during internal and external validation because there is (1) no need to update model weights during evaluation settings and (2) need for comparing models against a consistent benchmark and augmentation introduces randomness

Step 2: Online CXR feature extraction

Design decisions Two different previously published pre-trained DenseNet-121 CXR models [28] are tried for feature selection for our downstream mortality prediction task. The Madrid CXR images are resized during training to the input size for each of the pre-trained models (320×320 vs. 224×224) to output the imaging features for classification. The last fully connected layer of both models, containing 14 outputs corresponding to the 14 radiologic CheXpert finding labels [35], was removed. Instead, linearized convolutional features from either the second (–2) or the fourth (–4) to the last layer were used for the mortality prediction classification task. The pre-trained models were partially frozen, with model weights updating after either layer 355, 400, or 420 during training. Choices for which “teacher” pre-trained model, feature layer to use, and how many model layers to update for the new mortality prediction task are set up as hyperparameters to be tuned in our experiments

Reasons The Madrid dataset is too small to train deep learning networks from scratch. The pre-trained CXR models chosen have already been trained on much larger CXR datasets (MIMIC-CXR) [29] (>200,000 images) to discern features that are useful for diagnosing 14 different CXR lung and heart radiologic findings, which are also clinically relevant for COVID patients. The final few layers in pre-trained convolutional neural networks tend to have best summarized the features useful for downstream (related) classification tasks. Since we only have a small training dataset (Madrid), we decided to only partially update the weights in the later layers in the pre-trained models and leave the choice of how many layers to update as a tunable parameter—knowing that there is a balance to be “learned” between updating weights for the new task on the small Madrid training dataset and losing the benefit of pre-learned weights from the pre-trained “teacher” CXR models

Step 3: Mortality classification layers

Design decisions After CXR features are extracted from a pre-trained model, we added a classification block consisting of tunable number of hidden linear layers, followed by a final activation function (choice between ReLU and LeakyReLU), a dropout layer, and a single binary output layer. The output layer represents whether a patient is alive or expired at 30 days. An initial bias to the final out layer was optionally added and tuned along with the choices for activation function (ReLU or LeakyReLU) and the number and sizes of the hidden layers

Reasons The feature size extracted from both pre-trained models is 1024 in length. Additional classification layers were added to learn the new mortality classification task. Since the layer numbers and sizes are arbitrary, we picked a few common sizes to tune. We tried LeakyRelu as an activation function in the classification block because the CXR features extracted from the (–2) and (–4) layers can have many zeros due to the DenseNet-121 architecture. Adding initial bias to the output layer can help with performance for very unbalanced dataset

Training and model picking

All models**Step 4: Optimization settings**

<i>Design decisions</i>	Binary cross entropy was used as the loss function and the Adam optimizer was used for parameter optimization. We did not tune for these settings
<i>Reasons</i>	Binary cross entropy as the loss is appropriate for the binary mortality classification task. Adam is a fast optimizer, helps with avoiding overfitting and has shown good performance over a range of tasks

Step 5: Hyperparameter tuning and model selection

<i>Design decisions</i>	Supplementary Table 4 provides a summary of all the hyperparameters we experimented with on the Madrid dataset to select for the final best performing CXR-based COVID-19 30-day prediction model. An experiment is defined by one unique combination of hyperparameters. Due to limited training resources and a large hyperparameter search space (345,600 unique combinations), we had to first rough search and manually narrow down the hyperparameter search space—e.g., early observation suggests most experiments did better with smaller batch sizes, LeakyReLU activation, and with Bbox augmentation. We then fine-tuned the model on the other more important parameters such as the learning rate. Early stopping was used to end experiments that did not show loss reduction after 2 or 5 epochs. Overall, we performed over 300 experiments. For each experiment, we plotted the train and valid curves for multiple metrics (recall, precision, accuracy, AUC and F1-score) against the number of epochs. We performed a range of manual and automatic model selection by (1) evaluating experiments with F1-scores above 0.25 for all four folds and (2) manually examining the train-vs.-validation learning curves to pick the hyperparameter setting that showed improvement of the model's precision and recall from baseline for both the train and valid data, as well as ensuring that the chosen model did not show evidence of overfitting
<i>Reasons</i>	The standard practice for hyperparameter tuning is to update model weights on the train dataset and evaluate the updated model on the validation dataset at the end of each epoch, which is when the model has “seen” all examples in the train set once. Despite using all of the Madrid dataset for training and validation, the number of positive cases in the valid set is still small. Simply picking the best F1-score automatically without inspecting all the learning curves could just end up picking a “lucky” epoch

EHR-CXR fusion model

<i>Design decisions</i>	We took a late fusion approach that uses the output probability from the CXR model as a feature along with the EHR features for the 30-day mortality classification. With the Madrid train dataset, we again tuned four different machine learning models (logistic regression, random forest, gradient boosting, and XGBoost) in a fourfold cross-validation setting and the best model along with the best hyperparameters were selected using randomized grid search via the same methodology as that for training the EHR-based model
<i>Reasons</i>	Late fusion approach is used because it can be implemented with traditional machine learning methods, which can avoid overfitting for smaller datasets. On the other hand, intermediate (joint) fusion implemented by neural networks requires more data for training (the implementation of the intermediate fusion model can also be found in Supplementary Table 6). In addition, the much larger feature size from imaging modality can easily swamp important clinical signals from the tabular EHR data. From analyzing the fusion model's point of view, late fusion allows interpretation of the overall feature importance from the CXR model's prediction

Model evaluation

<i>Design decisions</i>	We made a clear separation between model developers and final model testers. Development of models includes programming feature selection and model training. External testing of models requires institutional access for the Hoboken and Seoul data, which were obtained upon request with submission of our study protocol
<i>Reasons</i>	This is the best practice to avoid repeated testing on the final test datasets, which could invalidate the reported results. It is also a common setting in real life model evaluation scenarios

Code packaging for testing

<i>Design decisions</i>	We packaged the inference code for the three different models for testing in an end-to-end Colab Notebook for the model testers to run on their datasets
<i>Reasons</i>	All datasets had been de-identified and are hosted on different HIPPA compliant cloud servers with access granted to different researchers based on institutional affiliation, data access approvals, and IRBs. Running via Colab, which have access management protocols, allows the clinical researchers to run the inference code without setting up Python and other required packages on their local machines, which can be a technical barrier

Table 3 Internal validation on Madrid dataset with 95% confidence intervals

	EHR-based	CXR-based	Fusion
AUROC (CI)	0.82 (0.79–0.84)	0.81 (0.78–0.83)	0.85 (0.83–0.87)
Sensitivity (CI)	0.77 (0.71–0.82)	0.76 (0.71–0.82)	0.79 (0.74–0.84)
Specificity (CI)	0.71 (0.66–0.76)	0.72 (0.67–0.75)	0.74 (0.71–0.78)
PPV (CI)	0.24 (0.21–0.28)	0.25 (0.21–0.28)	0.27 (0.23–0.31)
NPV (CI)	0.96 (0.95–0.97)	0.96 (0.95–0.97)	0.97 (0.96–0.98)
F1-score (CI)	0.36 (0.32–0.41)	0.37 (0.33–0.41)	0.40 (0.36–0.45)
Accuracy (CI)	0.71 (0.68–0.76)	0.73 (0.68–0.76)	0.75 (0.72–0.78)

AUROC area under the receiver operating characteristic curve, PPV positive predictive value, NPV negative predictive value, CI confidence interval

Data Sharing

We have made the code and other resources to reproduce our model training process available to the research community. The authors provided open access to all their data extraction, filtering, data wrangling, modeling, figures and tables, code, and queries on https://github.com/theonesp/multimodal_mortality_covid. The de-identified version of Madrid COVID Data Saves Lives repository can be requested at <https://www.hmhosptiales.com/coronavirus/covid-data-save-lives/english-version>. The associated datasets in this study can be accessed through the respective application processes of the hospitals involved.

Results

Internal Validation

Table 3 shows the results of internal validation for mortality prediction in COVID-19 patients. The models were trained and validated on the Madrid dataset using fourfold cross-validation. The respective F1-score (95% confidence interval) of the EHR-based, CXR-based, and fusion models were 0.36 (95% CI 0.32–0.41), 0.37 (0.33–0.41), and 0.40

(0.36–0.45). Figure 3(A) shows the ROC curves obtained from EHR-based (orange), CXR-based (green), and fusion models (blue) for internal validation on Madrid datasets.

External Testing

Table 4 shows the results of external testing for mortality prediction using all the Madrid dataset for model development and Hoboken and Seoul datasets for external testing. In the external testing on the Hoboken dataset, the F1-score (95% CI) of the EHR-based, CXR-based, and fusion models were 0.66 (0.59–0.73), 0.64 (0.57–0.70), and 0.69 (0.62–0.76), respectively. In the external testing on the Seoul dataset, the respective F1-score (95%CI) of the EHR-based, CXR-based, and fusion models were 0.15(0.04–0.28), 0.13 (0.03–0.25), and 0.21 (0.06–0.38). The ROC curves for external testing on Hoboken and Seoul datasets are illustrated in Fig. 3(B) and (C), respectively.

Explainability Analysis

Figures 4 and 5 show the impact of features on the EHR-based and fusion models' prediction, respectively. Figure 5

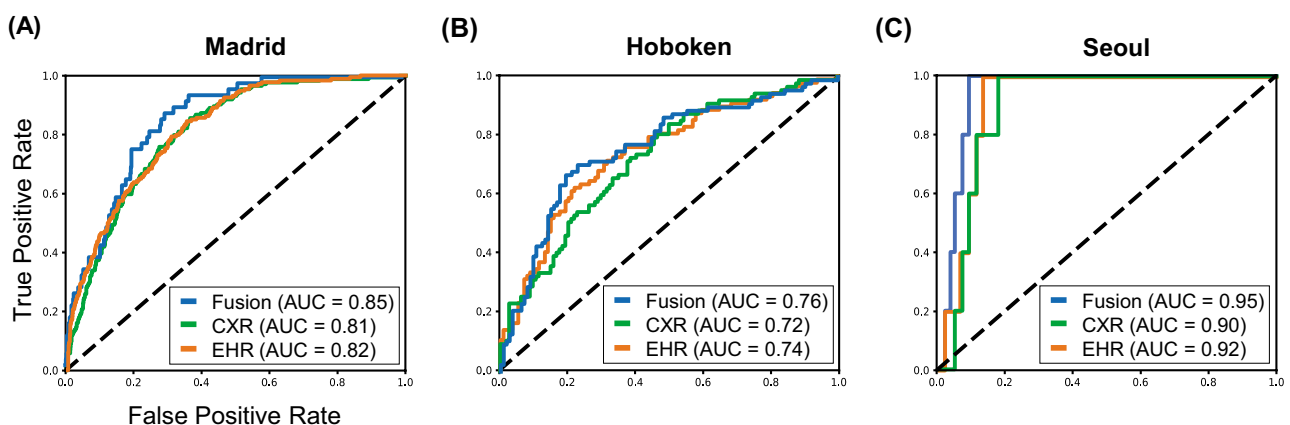


Fig. 3 Model performance using EHR-based model, CXR-based model, and fusion model (EHR+CXR). (A) Internal validation on Madrid dataset; (B) external testing on Hoboken dataset; and (C) external testing on Seoul dataset

Table 4 External testing on Hoboken and Seoul datasets with 95% confidence intervals

	Hoboken dataset			Seoul dataset		
	EHR-based	CXR-based	Fusion	EHR-based	CXR-based	Fusion
AUROC (CI)	0.74 (0.68–0.80)	0.72 (0.66–0.78)	0.76 (0.70–0.82)	0.92 (0.88–0.96)	0.90 (0.86–0.94)	0.95 (0.92–0.98)
Sensitivity (CI)	0.68 (0.59–0.77)	0.68 (0.57–0.8)	0.68 (0.60–0.76)	0.64 (0.25–0.86)	0.63 (0.20–0.86)	0.64 (0.20–0.86)
Specificity (CI)	0.72 (0.62–0.82)	0.65 (0.55–0.78)	0.78 (0.70–0.85)	0.88 (0.85–0.93)	0.86 (0.80–0.93)	0.93 (0.89–0.96)
PPV (CI)	0.65 (0.56–0.75)	0.60 (0.52–0.69)	0.71 (0.61–0.79)	0.09 (0.02–0.17)	0.07 (0.02–0.15)	0.13 (0.03–0.25)
NPV (CI)	0.75 (0.68–0.81)	0.73 (0.66–0.8)	0.76 (0.70–0.82)	0.99 (0.99–1.0)	0.99 (0.99–1.0)	1.00 (0.99–1.0)
F1-score (CI)	0.66 (0.59–0.73)	0.64 (0.57–0.7)	0.69 (0.62–0.76)	0.15 (0.04–0.28)	0.13 (0.03–0.25)	0.21 (0.06–0.38)
Accuracy (CI)	0.70 (0.65–0.76)	0.67 (0.61–0.72)	0.74 (0.68–0.79)	0.92 (0.88–0.96)	0.90 (0.86–0.94)	0.95 (0.92–0.98)

AUROC area under the receiver operating characteristic curve, PPV positive predictive value, NPV negative predictive value, CI confidence interval

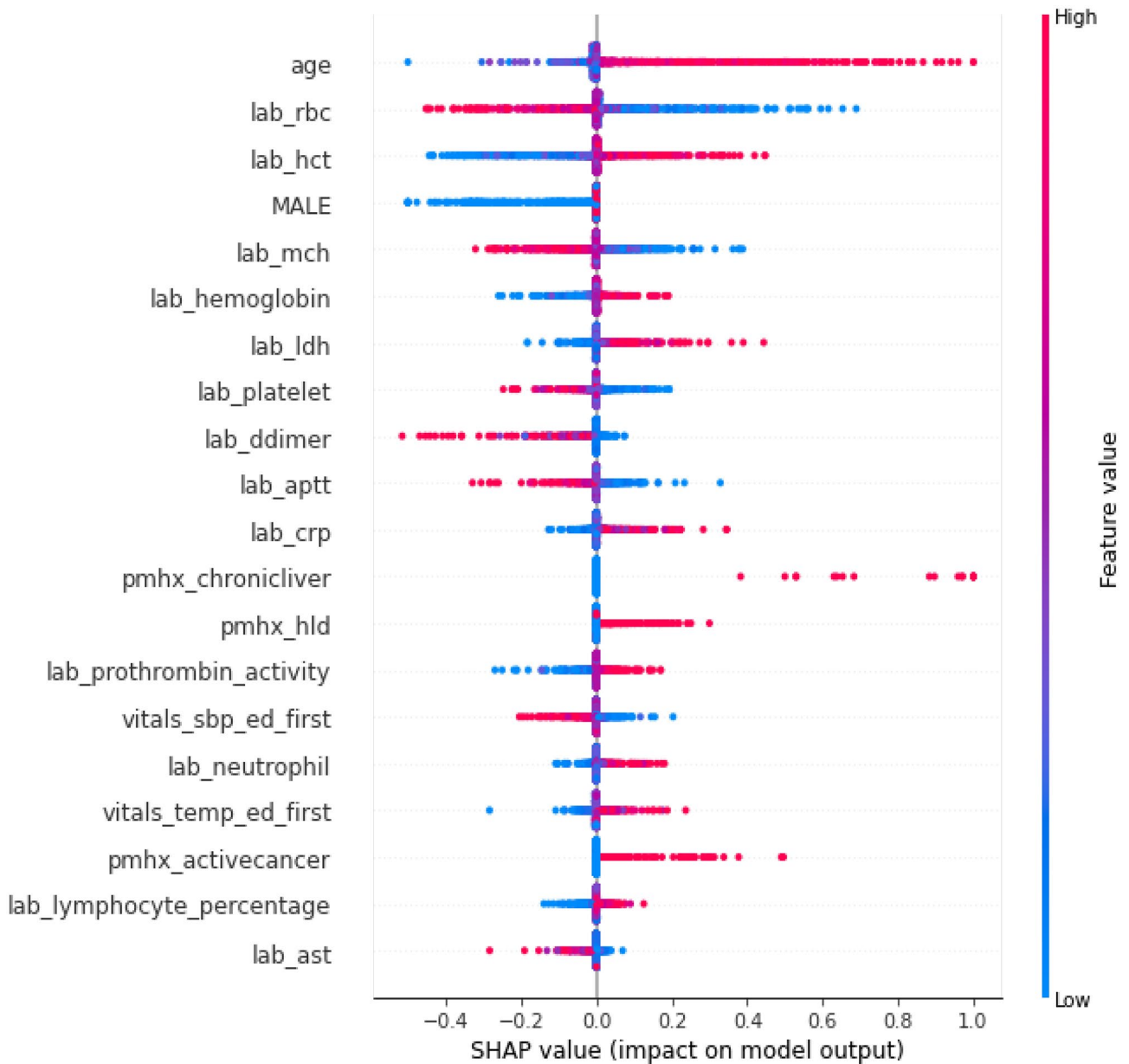


Fig. 4 Feature importance of the EHR-based model revealed by a SHAP plot. Features on the y-axis are ranked by their mean absolute SHAP values and each point represents a patient

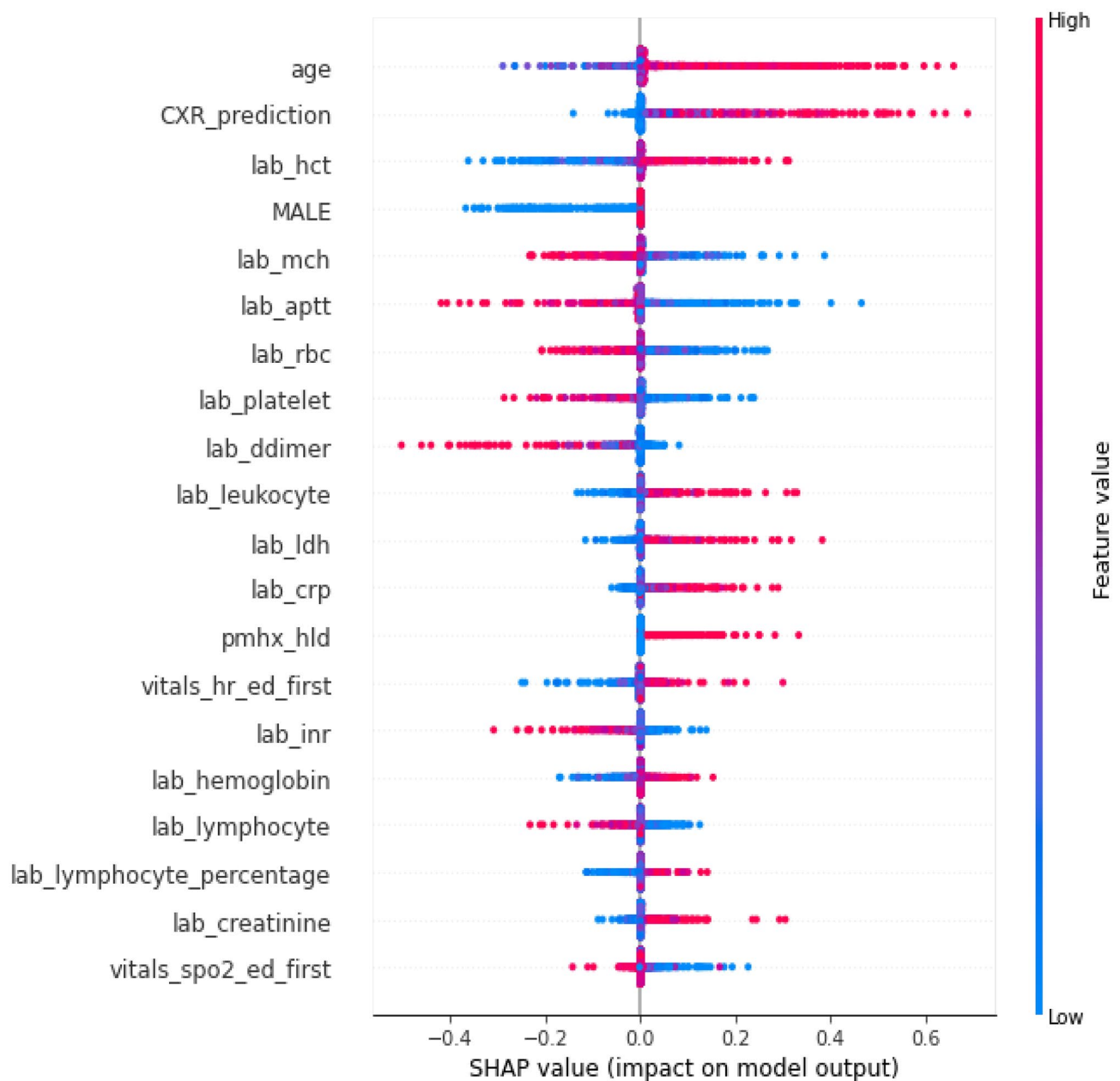


Fig. 5 Feature importance of the fusion model revealed by a SHAP plot. Features on the y-axis are ranked by their mean absolute SHAP values and each point represents a patient

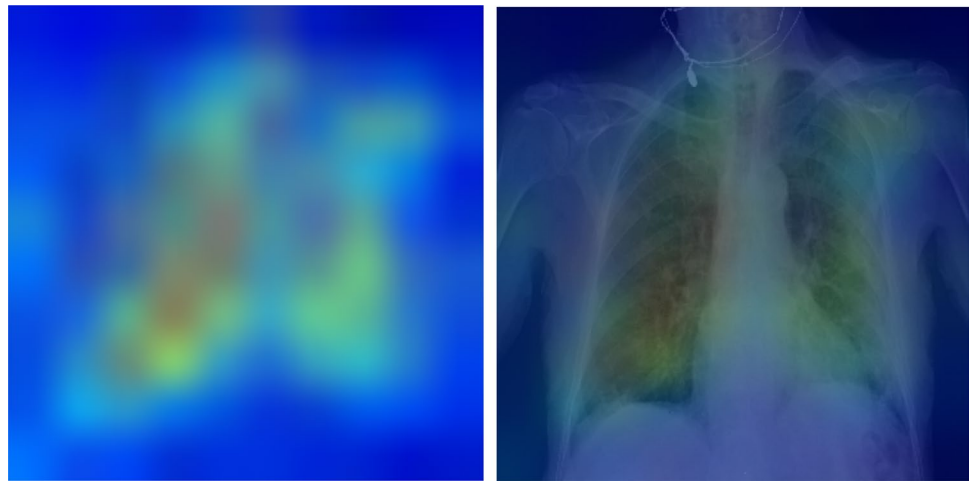
shows the CXR model's prediction and the patient age were two features with highest mean absolute SHapley Additive exPlanations (SHAP) [21] value for predicting 30-day mortality. Figure 6 shows the mean Gradient-weighted Class Activation Mapping (Grad-CAM) [22] heatmap obtained by averaging the heatmaps with prediction probability larger than 0.6 for the expired patients in the Madrid dataset. CXR regions with high levels of importance in the model prediction (represented in red or yellow) were located within the

lung zones. In particular, the lung parenchyma and mediastinal structures were the primary focus of the algorithm for mortality prediction.

Fairness Analysis

Supplementary Table 5 shows the difference of model performance between female vs. male patients across all three datasets. All the expired cases in the Seoul dataset are male.

Fig. 6 Explainability: heatmaps using Grad-CAM algorithm shows that the model primarily uses imaging features from the lungs and mediastinum region for mortality prediction. The image was produced by averaging the heatmaps from the expired patients with prediction probability larger than 0.6 and overlaying it on an actual CXR so it is easier to highlight the physiologic area



Discussion

Our findings demonstrate differences in the performance of our predictive models across different institutions, clinical settings, and populations. Previous studies have demonstrated that predictive models tend not to perform well outside the institution and setting that it was trained in, while also losing their accuracy over time due to underlying clinical data drift [30–32]. In this paper, we trained three models (EHR-based, CXR-based, and fusion) by optimizing their F1-scores for 30-day mortality prediction from hospital admission for confirmed COVID-19 patients. On internal validation (Madrid) and testing on two external datasets (Hoboken and Seoul), point estimates of the F1-score of the fusion model consistently outperformed the EHR-based and the CXR-based models. These findings are not statistically significant at 95% CI, which can be expected in the context of small numbers of mortality events in all the datasets. We reported all metrics for transparent reporting on our models' performance. Reporting just AUROC and/or accuracy can give a falsely higher sense of model performance for imbalanced datasets.

On evaluating the changes in F1-scores between the results on the Madrid dataset and the two external test sets, we see expected significantly drops in F1-scores for all three models when they are tested on the Seoul dataset (statistically significant). However, we see unexpectedly higher F1-scores from all three models for the Hoboken dataset (statistically significant). Possible explanations for these findings include (1) the surge in COVID-19 cases during this period in the Greater New York Metropolitan area, which includes Hoboken, NJ, that led to higher mortality rates in the hospital (more likely), and (2) the pre-trained “teacher” CXR models used in the development of our CXR model were trained on CXR images from the USA (possible). Both factors would make Hoboken an easier evaluation dataset than the both the Madrid and Seoul datasets. For the former factor, F1-scores

would naturally be higher if the target prevalence is higher in a dataset. For the latter factor, since deep learning models are brittle to small changes in machine type and calibration, changes in geography (a different country with likely more different machines and calibration protocols) alone may affect model's feature extraction suitability for the same task—i.e., it is possible that the CXR models extract better features from Hoboken CXRs as the images were taken in similar US setting.

Our findings further highlight the need for a guided strategy with the use of predictive models as clinical decision support tools. This is particularly important for machine learning models because, unlike traditional statistically based models (e.g., multivariate regression models which have deterministic performance on the same data), the state-of-the-art machine learning models, though powerful, are often built with many design decisions during their development. Changes in any parameters, optimizing target(s) (e.g., accuracy, AUROC, and F1-score), or even just different GPUs or TPUs for training, could result in different performance. As such, performance reported in any paper is really a snapshot. Furthermore, institutional and temporal data variation can also change the performance of these models. Ultimately, performance of predictive models, like most medical tests, is sensitive to changes in disease severity, prevalence, and distribution in a patient population.

Therefore, we reinforce the previously reported recommendation that institutions should reassess models on local datasets [15, 33]. They should also consider fine tuning their own predictive models and to learn what works best for their local patient populations. Training a predictive model to prioritize a high positive predictive value may mitigate the risk of incorrectly predicting an outcome—in this case, mortality—among patients who would have otherwise survived or have not met the outcome. Prioritization of sensitivity may allow clinicians to triage as many

severely ill patients to advanced care facilities. Ultimately, locally tested predictive models can become tools that help institutions address their individual needs, either to appropriately distribute limited resources or to help detect and manage severely ill patients early. To facilitate this and for reproducibility, we have open sourced our data preprocessing and training code for re-use by other research groups and institutions.

Multiple groups have suggested best practices and regulations in the development and use of artificial intelligence and predictive modeling that address pertinent concerns [15, 33]. Much of these recommendations have to do with reproducibility, quality of data being used, and the intended function of artificial intelligence programs [34]. However, the complexity of applying machine learning models in clinical settings goes beyond reproducibility and generalizability. Deep learning's advantage of not needing to engineer predictive features for model building can be offset by the disadvantage of its lack of explainability. Models may draw spurious associations between confounding tabular or imaging features and the outcome of interest [12, 16]. For example, a prior study has shown that AI has a predilection for detecting imaging features other than signals of pathology as shortcuts for predicting COVID-19 outcomes [12]. Unlike linear models, weights in deep learning models have no intrinsic significance on their own that can be interpreted clinically outside the model. Despite this, in the imaging space, researchers have used methods, such as Grad-CAM, to post analyze the "explainability" of deep learning imaging models by examining where on the image the trained model "attended" to most for prediction. However, these analyses are often qualitative in nature for publication purposes, which we argue is insufficiently rigorous for most clinical applications.

Therefore, for imaging feature explainability, we not only presented heatmaps from post-training Grad-CAM analysis of the CXR model but also specifically used anatomical bounding box augmentation to teach our CXR model to focus on lungs and mediastinum regions for prediction during the training stage. Choices for both tuning with or without anatomy augmentation was used and the best model had utilized the anatomical regional augmentation approach. As shown in Grad-CAM analysis, our CXR model does focus on the clinically important lungs and mediastinum CXR regions for prediction. This gives our clinicians more confidence for the model's prediction.

Lastly, we took a late fusion approach to model features from the EHR data and the CXR image so that we could easily assess how much the prediction depended on different data sources via SHAP plots. Previous machine learning models have also used similar techniques to elucidate the explainability of their models [12, 36]. In our case, the fusion model placed the most weight on patient age and

CXR features, further supporting that including both clinical and imaging data improves downstream model performance.

This study is limited by its retrospective nature, imbalanced and small evaluation datasets, and merits further research. The models trained also need the exact same input features at inference and some institutions may not routinely collect all the required input variables. In addition, as with many published AI models, fairness as an operationalized outcome has not been incorporated in our models [33]. We did, however, assess for differences in the models' performance separately for male and female on the Madrid and the Hoboken dataset (no female deaths in the Seoul dataset), which showed no statistically significant differences on 95% CI analysis except the EHR-based model on the Madrid dataset. In general, evaluating differences in model performance in subpopulations can help elucidate and inform downstream applications about potential problems if an AI model was applied to patients from under-represented/marginalized populations. Pooled results from the general population may gloss over worse outcomes in vulnerable groups [37]. Furthermore, in general when the dataset contains underlying data entry biases and/or imbalanced representations, building fair models is still an unsolved technical research problem.

Conclusion

Machine learning methods offer the advantage of utilizing richer clinical data for predictive modeling, which many have explored during the COVID-19 pandemic. However, many studies published so far have further exposed the credibility crisis that machine learning is facing in terms of reproducibility, generalizability, explainability, and fairness. This is often due to implementation issues, such as poorly documented study designs, lack of external test sets, and study code availability. In this paper, we employed best machine learning practices and trained three machine learning models on a model development (internal) dataset. We subsequently stress tested the final models on two external datasets from different countries. We redemonstrated (1) that using features from both the EHR and CXR imaging data resulted in better 30-day mortality prediction performances across all three datasets, and (2) the need to fine tune models on local datasets and update with time. We evaluated our models for explainability in terms of feature dependence, and fairness in terms of gender-based performance differences. Finally, for the sake of transparency and reproducibility, we documented all study design decisions and made the study code available to the research community.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10278-022-00674-z>.

Acknowledgements There is no funding for this project. PK is funded by the Ministry of Science and Technology, Taiwan (MOST109-2222-E-007-004-MY3). LAC is funded by the National Institute of Health through the NIBIB R01 EB017205. ECD is funded in part through the Cancer Center Support Grant from the National Cancer Institute (P30 CA008748).

Author contributions JTW, MA, PK, WY, JAP, JSY, JMC, and LAC conceived and designed the study. MA, JAP, and JSY collected data. JAP, JSY, and ECD reviewed the CXR images. JTW, MA, PK, and WY implemented the models and did the data analysis. JTW, MA, PK, JAP, JSY, and LAC drafted the manuscript. All authors revised, reviewed, and approved the manuscript. Leo Anthony Cel José Maria Castellano are co-senior authors.

Funding Open Access funding provided by the MIT Libraries

Declarations

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.


References

1. M. Xu *et al.*, “Accurately Differentiating COVID-19, Other Viral Infection, and Healthy Individuals Using Multimodal Features via Late Fusion Learning,” *medRxiv*, p. 2020.08.18.20176776, Aug. 2020, <https://doi.org/10.1101/2020.08.18.20176776>.
2. G. Chassagnon and N. Paragios, “Holistic AI-Driven Quantification, Staging and Prognosis of COVID-19 Pneumonia,” *medRxiv*, p. 2020.04.17.20069187, Jul. 2020, <https://doi.org/10.1101/2020.04.17.20069187>.
3. X. Wang *et al.*, “Multicenter Study of Temporal Changes and Prognostic Value of a CT Visual Severity Score in Hospitalized Patients With Coronavirus Disease (COVID-19),” *Am. J. Roentgenol.*, pp. 1–10, Sep. 2020, <https://doi.org/10.2214/AJR.20.24044>.
4. T. Ramtohl *et al.*, “Quantitative CT Extent of Lung Damage in COVID-19 Pneumonia Is an Independent Risk Factor for Inpatient Mortality in a Population of Cancer Patients: A Prospective Study,” *Front. Oncol.*, vol. 10, Sep. 2020, <https://doi.org/10.3389/fonc.2020.01560>.
5. N. Lassau *et al.*, “Integration of clinical characteristics, lab tests and a deep learning CT scan analysis to predict severity of hospitalized COVID-19 patients,” *medRxiv*, p. 2020.05.14.20101972, Oct. 2020, <https://doi.org/10.1101/2020.05.14.20101972>.
6. Q. Wu *et al.*, “Radiomics Analysis of Computed Tomography helps predict poor prognostic outcome in COVID-19,” *Theranostics*, vol. 10, no. 16, pp. 7231–7244, 2020, <https://doi.org/10.7150/thno.46428>.
7. Y. Zheng *et al.*, “Development and Validation of a Prognostic Nomogram Based on Clinical and CT Features for Adverse Outcome Prediction in Patients with COVID-19,” *Korean J. Radiol.*, vol. 21, no. 8, pp. 1007–1017, Aug. 2020, <https://doi.org/10.3348/kjr.2020.0485>.
8. H. Chao *et al.*, “Integrative analysis for COVID-19 patient outcome prediction,” *Med. Image Anal.*, vol. 67, p. 101844, Jan. 2021, <https://doi.org/10.1016/j.media.2020.101844>.
9. M. Roberts *et al.*, “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans,” *Nat. Mach. Intell.*, vol. 3, no. 3, Art. no. 3, Mar. 2021, <https://doi.org/10.1038/s42256-021-00307-0>.
10. J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi, “The myth of generalisability in clinical research and machine learning in health care,” *Lancet Digit. Health*, vol. 2, no. 9, pp. e489–e492, Sep. 2020, [https://doi.org/10.1016/S2589-7500\(20\)30186-2](https://doi.org/10.1016/S2589-7500(20)30186-2).
11. S. on Facebook, S. on Twitter, and S. on LinkedIn, “Major flaws found in machine learning for COVID-19 diagnosis,” *VentureBeat*, Mar. 23, 2021. <https://venturebeat.com/2021/03/23/major-flaws-found-in-machine-learning-for-covid-19-diagnosis/> (accessed Jun. 04, 2021).
12. A. J. DeGrave, J. D. Janizek, and S.-I. Lee, “AI for radiographic COVID-19 detection selects shortcuts over signal,” *Nat. Mach. Intell.*, pp. 1–10, May 2021, <https://doi.org/10.1038/s42256-021-00338-7>.
13. U. J. Muehlematter, P. Daniore, and K. N. Vokinger, “Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis,” *Lancet Digit. Health*, vol. 3, no. 3, pp. e195–e203, Mar. 2021, [https://doi.org/10.1016/S2589-7500\(20\)30292-2](https://doi.org/10.1016/S2589-7500(20)30292-2).
14. S. Benjamins, P. Dhunoo, and B. Meskó, “The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database,” *Npj Digit. Med.*, vol. 3, no. 1, Art. no. 1, Sep. 2020, <https://doi.org/10.1038/s41746-020-00324-0>.
15. E. Wu, K. Wu, R. Daneshjou, D. Ouyang, D. E. Ho, and J. Zou, “How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals,” *Nat. Med.*, vol. 27, no. 4, Art. no. 4, Apr. 2021, <https://doi.org/10.1038/s41591-021-01312-x>.
16. L. Wynants *et al.*, “Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal,” *BMJ*, vol. 369, p. m1328, Apr. 2020, <https://doi.org/10.1136/bmj.m1328>.
17. J. Mongan, L. Moy, and C. E. Kahn, “Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers,” *Radiol. Artif. Intell.*, vol. 2, no. 2, p. e200029, Mar. 2020, <https://doi.org/10.1148/ryai.2020200029>.
18. “Covid Data Save Lives English Version.” <https://www.hmhosptales.com/coronavirus/covid-data-save-lives/english-version> (accessed Jun. 24, 2021).
19. J. S. Yao *et al.*, “The Minimal Effect of Zinc on the Survival of Hospitalized Patients With COVID-19: An Observational Study,” *Chest*, vol. 159, no. 1, pp. 108–111, Jan. 2021, <https://doi.org/10.1016/j.chest.2020.06.082>.
20. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
21. S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *ArXiv170507874 Cs Stat*, Nov. 2017, Accessed: Jun. 27, 2021. [Online]. Available: <http://arxiv.org/abs/1705.07874>
22. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020, <https://doi.org/10.1007/s11263-019-01228-7>.

23. O. A. Panagiotou *et al.*, “Risk Factors Associated With All-Cause 30-Day Mortality in Nursing Home Residents With COVID-19,” *JAMA Intern. Med.*, vol. 181, no. 4, p. 439, Apr. 2021, <https://doi.org/10.1001/jamainternmed.2020.7968>.
24. Y. X. Gue, M. Tennyson, J. Gao, S. Ren, R. Kanji, and D. A. Gorog, “Development of a novel risk score to predict mortality in patients admitted to hospital with COVID-19,” *Sci. Rep.*, vol. 10, no. 1, Art. no. 1, Dec. 2020, <https://doi.org/10.1038/s41598-020-78505-w>.
25. J. Berenguer *et al.*, “Development and validation of a prediction model for 30-day mortality in hospitalised patients with COVID-19: the COVID-19 SEIMC score,” *Thorax*, Feb. 2021, <https://doi.org/10.1136/thoraxjnl-2020-216001>.
26. D. A. Asch *et al.*, “Variation in US Hospital Mortality Rates for Patients Admitted With COVID-19 During the First 6 Months of the Pandemic,” *JAMA Intern. Med.*, vol. 181, no. 4, pp. 471–478, Apr. 2021, <https://doi.org/10.1001/jamainternmed.2020.8193>.
27. J. Wu *et al.*, “Automatic Bounding Box Annotation of Chest X-Ray Data for Localization of Abnormalities,” in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Iowa City, IA, USA, Apr. 2020, pp. 799–803. <https://doi.org/10.1109/ISBI45749.2020.9098482>.
28. J. Irvin *et al.*, “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison,” *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, Art. no. 01, Jul. 2019, <https://doi.org/10.1609/aaai.v33i01.3301590>.
29. A. E. W. Johnson *et al.*, “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” *Sci. Data*, vol. 6, no. 1, p. 317, Dec. 2019, <https://doi.org/10.1038/s41597-019-0322-0>.
30. S. E. Davis, T. A. Lasko, G. Chen, and M. E. Matheny, “Calibration Drift Among Regression and Machine Learning Models for Hospital Mortality,” *AMIA Annu. Symp. Proc. AMIA Symp.*, vol. 2017, pp. 625–634, 2017.
31. S. E. Davis, T. A. Lasko, G. Chen, E. D. Siew, and M. E. Matheny, “Calibration drift in regression and machine learning models for acute kidney injury,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 24, no. 6, pp. 1052–1061, Nov. 2017, <https://doi.org/10.1093/jamia/ocx030>.
32. E. Paul, M. Bailey, A. Van Lint, and V. Pilcher, “Performance of APACHE III over time in Australia and New Zealand: a retrospective cohort study,” *Anaesth. Intensive Care*, vol. 40, no. 6, pp. 980–994, Nov. 2012, <https://doi.org/10.1177/0310057X1204000609>.
33. J. Wawira Gichoya, L. G. McCoy, L. A. Celi, and M. Ghassemi, “Equity in essence: a call for operationalising fairness in machine learning for healthcare,” *BMJ Health Amp Care Inform.*, vol. 28, no. 1, p. e100289, Apr. 2021, <https://doi.org/10.1136/bmjhci-2020-100289>.
34. [34]S. E. Davis, R. A. Greevy Jr, C. Fonnesbeck, T. A. Lasko, C. G. Walsh, and M. E. Matheny, “A nonparametric updating method to correct clinical prediction model drift,” *J. Am. Med. Inform. Assoc.*, vol. 26, no. 12, pp. 1448–1457, 2019.
35. P.-C. Kuo *et al.*, “Recalibration of deep learning models for abnormality detection in smartphone-captured chest radiograph,” *NPJ Digit. Med.*, vol. 4, no. 1, p. 25, Feb. 2021, <https://doi.org/10.1038/s41746-021-00393-9>.
36. D. Bertsimas *et al.*, “COVID-19 mortality risk assessment: An international multi-center study,” *PLOS ONE*, vol. 15, no. 12, p. e0243262, Dec. 2020, <https://doi.org/10.1371/journal.pone.0243262>.
37. R. Benjamin, “Assessing risk, automating racism,” *Science*, vol. 366, no. 6464, p. 421, Oct. 2019, <https://doi.org/10.1126/science.aaz3873>.

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

Authors and Affiliations

Joy Tzung-yu Wu¹ · Miguel Ángel Armengol de la Hoz^{2,3,4} · Po-Chih Kuo^{2,5}  · Joseph Alexander Paguio^{6,9} · Jasper Seth Yao^{6,9} · Edward Christopher Dee⁷ · Wesley Yeung^{2,8} · Jerry Jurado⁹ · Achintya Moulick⁹ · Carmelo Milazzo⁹ · Paloma Peinado¹⁰ · Paula Villares¹⁰ · Antonio Cubillo¹⁰ · José Felipe Varona¹⁰ · Hyung-Chul Lee¹¹ · Alberto Estirado¹⁰ · José Maria Castellano^{10,12} · Leo Anthony Celi^{2,13,14}

¹ Department of Radiology and Nuclear Medicine, Stanford University, Palo Alto, CA, USA

² Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA

³ Department of Anesthesia, Critical Care and Pain Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

⁴ Big Data Department, Fundacion Progreso Y Salud, Regional Ministry of Health of Andalucía, Andalucía, Spain

⁵ Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

⁶ Albert Einstein Medical Center, Philadelphia, PA, USA

⁷ Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

⁸ National University Heart Center, National University Hospital, Singapore, Singapore

⁹ Hoboken University Medical Center–CarePoint Health, Hoboken, NJ, USA

¹⁰ Centro Integral de Enfermedades Cardiovasculares, Hospital Universitario Montepíncipe, Grupo HM Hospitales, Madrid, Spain

¹¹ Department of Anesthesiology and Pain Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea

¹² Centro Nacional de Investigaciones Cardiovasculares, Instituto de Salud Carlos III, Madrid, Spain

¹³ Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

¹⁴ Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA