European Radiology
https://doi.org/10.1007/s00330-022-08588-8

CHEST

COVID-19 mortality prediction in the intensive care unit with deep learning based on longitudinal chest X-rays and clinical data

Jianhong Cheng[1] · John Sollee[2,3] · Celina Hsieh[2,3] · Hailin Yue[1] · Nicholas Vandal[4] · Justin Shanahan[4] · Ji Whae Choi[2,3] · Thi My Linh Tran[2,3] · Kasey Halsey[2,3] · Franklin Iheanacho[2,3] · James Warren[5] · Abdullah Ahmed[2,3] · Carsten Eickhoff[6] · Michael Feldman[7] · Eduardo Mortani Barbosa Jr[4] · Ihab Kamel[8] · Cheng Ting Lin[8] · Thomas Yi[2,3] · Terrance Healey[2,3] · Paul Zhang[4] · Jing Wu[1] · Michael Atalay[2,3] · Harrison X. Bai[8] · Zhicheng Jiao[2,3] · Jianxin Wang[1]

Received: 28 October 2021 / Revised: 14 December 2021 / Accepted: 22 January 2022

Abstract

Objectives We aimed to develop deep learning models using longitudinal chest X-rays (CXRs) and clinical data to predict in-hospital mortality of COVID-19 patients in the intensive care unit (ICU).

Methods Six hundred fifty-four patients (212 deceased, 442 alive, 5645 total CXRs) were identified across two institutions. Imaging and clinical data from one institution were used to train five longitudinal transformer-based networks applying five-fold cross-validation. The models were tested on data from the other institution, and pairwise comparisons were used to determine the best-performing models.

Results A higher proportion of deceased patients had elevated white blood cell count, decreased absolute lymphocyte count, elevated creatine concentration, and incidence of cardiovascular and chronic kidney disease. A model based on pre-ICU CXRs achieved an AUC of 0.632 and an accuracy of 0.593, and a model based on ICU CXRs achieved an AUC of 0.697 and an accuracy of 0.657. A model based on all longitudinal CXRs (both pre-ICU and ICU) achieved an AUC of 0.702 and an accuracy of 0.694. A model based on clinical data alone achieved an AUC of 0.653 and an accuracy of 0.657. The addition of longitudinal imaging to clinical data in a combined model significantly improved performance, reaching an AUC of 0.727 ($p = 0.039$) and an accuracy of 0.732.

Conclusions The addition of longitudinal CXRs to clinical data significantly improves mortality prediction with deep learning for COVID-19 patients in the ICU.

Key Points

- *Deep learning was used to predict mortality in COVID-19 ICU patients.*
- *Serial radiographs and clinical data were used.*
- *The models could inform clinical decision-making and resource allocation.*

Jianhong Cheng, John Sollee and Celina Hsieh contributed equally to this work.

✉ Harrison X. Bai
hbai7@jh.edu

✉ Zhicheng Jiao
Zhicheng_Jiao@Brown.edu

✉ Jianxin Wang
jxwang@mail.csu.edu.cn

[1] School of Computer Science and Engineering, Central South University of Technology, 932 Lushan S Rd, Yuelu District, Changsha, Hunan, China

[2] Department of Diagnostic Radiology, Rhode Island Hospital, 593 Eddy St., Providence, RI 02903, USA

[3] Warren Alpert Medical School of Brown University, Providence, RI 02903, USA

[4] Department of Diagnostic Radiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA

[5] Department of Data Science, University of London, London, UK

[6] Center for Biomedical Informatics, Brown University, Providence, RI 02912, USA

[7] Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA

[8] Department of Radiology and Radiological Sciences, Johns Hopkins University School of Medicine, 601 N Caroline St, Baltimore, MD 21205, USA

Published online: 19 February 2022

Keywords Artificial intelligence · Machine learning · Prognosis · Hospital mortality · Coronavirus

Abbreviations

AI	Artificial intelligence
AUC	Area under the receiver operating characteristic curve
COPD	Chronic obstructive pulmonary disease
CRP	C-reactive protein
CT	Computerized tomography
CVD	Cardiovascular disease
CXR	Chest X-ray
ED	Emergency department
HIV	Human immunodeficiency virus
HTN	Hypertension
ICU	Intensive care unit
IQR	Interquartile range
LTBN	Longitudinal transformer-based network
RF	Random forest
RT-PCR	Reverse transcriptase–polymerase chain reaction
SARS-CoV-2/COVID-19	Severe acute respiratory syndrome coronavirus 2 disease
SpO2	Oxygen saturation
ViT	Vision transformer
WBC	White blood cell count

Introduction

The novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) disease (COVID-19) was first detected in Wuhan, China, in late December 2019 and quickly became a global health crisis [1]. As of August 2021, almost 200 million confirmed cases have been reported globally with over four million deaths. In the USA alone, over 600,000 deaths are attributable to the virus [2]. Typical symptoms include fever, dyspnea, cough, and muscle aches; however, the disease can cause severe cardiorespiratory complications, particularly in vulnerable populations (e.g., the elderly and those with comorbidities) [3]. Despite rapid vaccine development and extensive public health mitigation efforts, COVID-19 remains a global health emergency. Additionally, novel variants threaten to exacerbate the severity and duration of the pandemic [4, 5].

Thoracic imaging, such as computerized tomography (CT) and chest radiograph (CXR), plays a key role not only in initial COVID-19 detection and diagnosis, but also in the continuous monitoring of disease progression and treatment efficacy during extended hospital stays [6–8]. While CXR is less sensitive for the detection of pneumonia associated with COVID-19 [9, 10]—particularly in less advanced stages—it is a helpful and versatile tool for monitoring the rapid pulmonary progression that is often seen in patients in the intensive care unit (ICU) [8, 11]. Moreover, longitudinal CXRs may provide vital information for risk stratification, clinical decision-making, and resource allocation [8]. Chest X-ray can be performed at the bedside in many cases, making it readily accessible and further increasing clinical utility, particularly in resource-limited settings [12].

Despite the potential of regular monitoring with CXRs to improve clinical care, longitudinal imaging is burdensome for radiologists. Given the current prevalence of COVID-19, manual, timely, and accurate interpretation of images is often logistically impossible, particularly for rapidly deteriorating ICU patients. Additionally, human readers are prone to variability, fatigue, and unconscious bias. To address these challenges, researchers have proposed artificial intelligence (AI) based tools to automate chest imaging interpretation and improve accuracy [11, 13–16]. For instance, AI with deep learning can predict the severity and progression of COVID-19 patients based on initial CXRs and clinical variables at presentation to the emergency department (ED) [16]. A model based on longitudinal CXRs may improve outcome prediction and inform clinical decision-making and resource allocation for critically ill patients. The purpose of this study was to develop deep learning models using longitudinal CXRs and clinical variables to predict in-hospital mortality of COVID-19 patients in the ICU.

Materials and methods

Clinical data acquisition and preprocessing

A retrospective chart review was performed between March 2020 and December 2020 to identify consecutive patients who presented to the EDs of two independent hospital systems, the University of Pennsylvania Health System in Philadelphia, PA, USA, and Brown University–affiliated hospitals in Providence, RI, USA. The institutional review boards of both institutions approved the study, and the requirement for written informed consent was waived. Patients were only included in the study if there was a positive reverse transcriptase–polymerase chain reaction (RT-PCR) test for COVID-19 (COVID-19 RT-PCR test; LabCorp). Furthermore, to focus outcome prediction on critically ill patients, only those who were admitted to the ICU were included. To allow for longitudinal assessment, only patients with at least two CXRs in anteroposterior view obtained in the ICU were included.

A subset of the data has previously been published [16–19]. In the study by Jiao et al [16], all patients from the University of Pennsylvania (N = 1834) and Brown University–affiliated hospitals (N = 475) who presented to the ED with a PCR-confirmed COVID-19 diagnosis were included. Deep learning was then used to predict disease severity and progression based on single-timepoint baseline chest X-rays and clinical variables. In the study by Wang et al [17], a further subset of the patients from the University of Pennsylvania (N = 144) and Brown University–affiliated hospitals (N = 31) who presented to the ED with a PCR-confirmed COVID-19 diagnosis and available baseline CT scans were included. Deep learning was then used to predict deterioration to critical illness based on imaging and clinical data. Two earlier studies used subsets of the patients in Wang et al [17] to assess the performance of radiologists in diagnosing COVID-19 [18] and the utility of AI to augment diagnosis by radiologists with baseline CT scans [19]. The current study expands upon the previous by predicting mortality in a subset of critically ill ICU patients from Jiao et al [16] based on longitudinal CXRs and clinical data.

For each patient, demographic, clinical, and laboratory variables taken on admission to the ICU including age, sex, temperature, oxygen saturation on room air (SpO2), absolute white blood cell count (WBC), absolute lymphocyte count, serum creatinine concentration, serum c-reactive protein (CRP) concentration, and comorbidities such as cardiovascular disease (CVD), hypertension (HTN), chronic obstructive pulmonary disease (COPD), chronic liver disease, chronic kidney disease, cancer, and human immunodeficiency virus (HIV) were collected. All continuous lab variables were binarized prior to analysis: fever was defined as a temperature of > 37 °C, low SpO2 as < 94%, high absolute WBC as > 11×10^9 cells/L, low absolute lymphocyte count as < 1×10^9 cells/L, high serum creatinine concentration as > 1.27 mg/dL, and high serum CRP concentration as > 1 mg/dL. The binary outcome of in-hospital mortality was also recorded.

Chest X-ray data acquisition and preprocessing

For patients meeting inclusion criteria, all CXRs obtained during ICU stay were identified (ICU CXRs). Furthermore, all CXRs obtained prior to ICU admission but during the hospital stay for the same disease course were identified (pre-ICU CXRs). CXRs with overall poor quality were excluded. Images were downloaded from the hospital picture archiving and communications system. Images were inverted as necessary so that air cavities appeared dark and padded and resized to 512 × 512 resolution. Then, pixel values were normalized and scaled to 0, 1. Finally, CXRs were segmented to generate lung masks for input to the deep learning model [16].

Deep learning architecture and training

Imaging and clinical variables from the University of Pennsylvania were used to train longitudinal transformer-based network (LTBN) models to predict the binary outcome of in-hospital mortality of COVID-19 patients in the ICU (Figure 1). The data from the University of Pennsylvania were randomly divided into two parts, 80% of which were used for training and 20% for internal validation. Finally, the models were tested on an external dataset derived from Brown University–affiliated hospitals. Five models were evaluated: (1) longitudinal CXRs before admission to the ICU ("pre-ICU model"), (2) longitudinal CXRs during the ICU stay ("ICU model"), (3) all longitudinal CXRs (pre-ICU and ICU) ("longitudinal model"), (4) demographic, clinical, and laboratory variables at the time of ICU admission only ("clinical model"), and (5) all longitudinal CXRs (pre-ICU and ICU) and clinical variables ("combined model").

For mortality prediction based on clinical variables, a model with three fully connected layers with 128, 32, and two neurons was established. To prevent overfitting, a dropout layer, which randomly set the input neuron to zero with a probability of 0.2, was embedded between the first two fully connected layers. To determine the relative importance of different clinical variables in predicting mortality, random forest (RF) models were utilized [20]. For mortality prediction based on CXRs, a LTBN consisting of Resnet-50 [21] and Vision Transformer (ViT) [22], termed "R50-ViT," was designed to extract both local and longitudinal global representation features. The proposed framework takes a series of longitudinal CXRs and the corresponding lung mask as input and generates features from the lung parenchyma region. The extracted features from all longitudinal CXRs are then combined using global average pooling and global max pooling operations. Finally, the combined features are fed into two fully connected layers with 1536 and two neurons and a softmax activation function to generate a probability score for mortality risk. The combined mortality prediction model was derived from the weighted sum of the longitudinal model and the clinical model, and the weights were obtained by training a fully connected layer. Additional details of the model architecture are provided in the supplementary materials (Figure R1).

The proposed models were implemented using Python (Version 3.6) and were run on two NVIDIA V100 GPUs for data parallel training. The network was trained with the Adam optimizer with an initial learning rate of 0.0005 and a poly learning rate strategy, in which the initial rate decays by each iteration with a power of 0.9. The batch size was set as one for each GPU, and the model was trained for 500 epochs. The codebase used in this study is available online (https://github.com/chengjianhong/Covid-19-CXR.git). The full dataset used to train and evaluate the models is not available for public access because of patient privacy concerns but is available

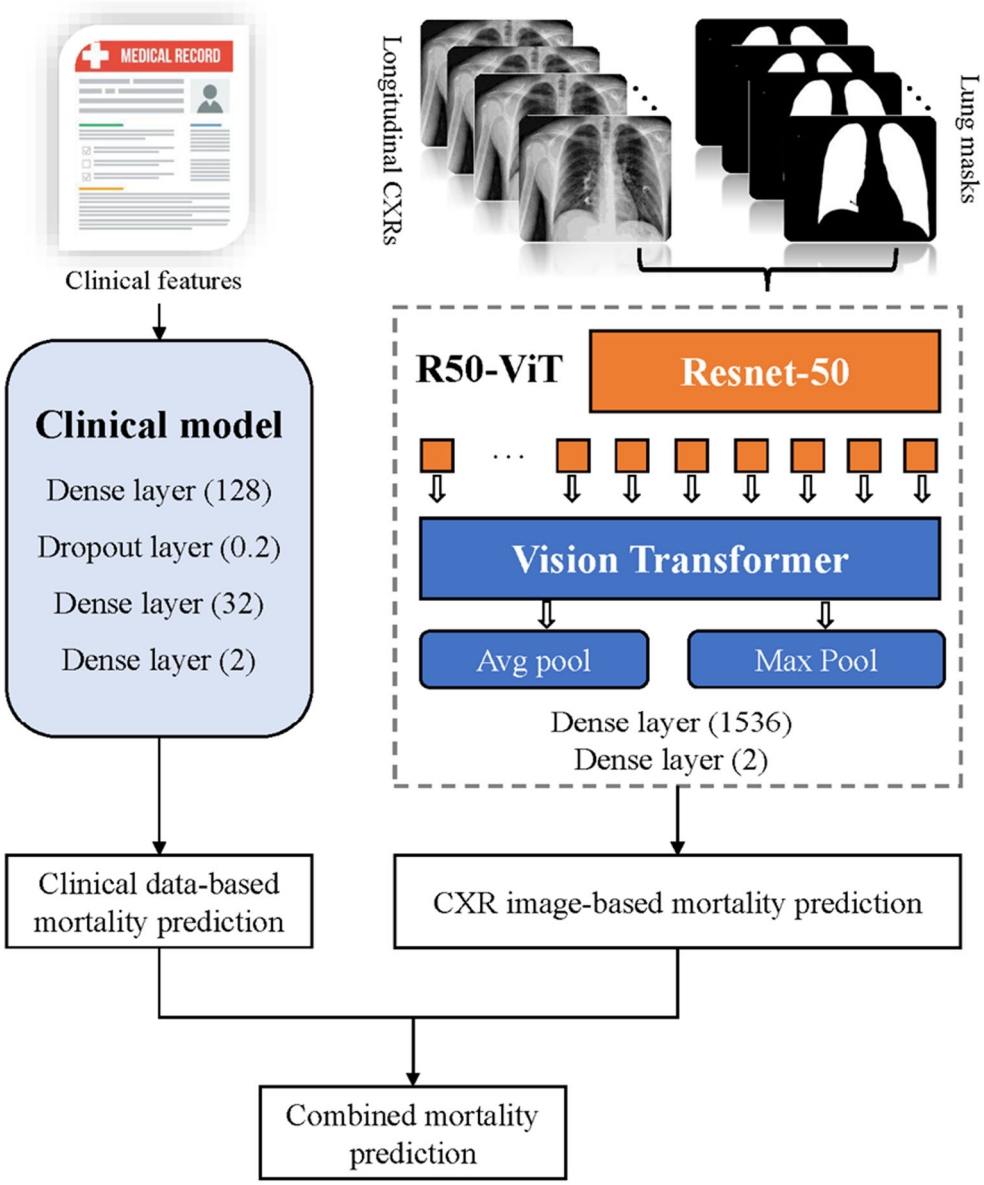


Fig. 1 Data analysis workflow and machine learning architecture. The clinical model consisted of three fully connected layers with 128, 32, and two neurons. A dropout layer with a probability of 0.2 was embedded between the first two layers. The CXR model used R50-ViT with two dense layers of 1536 and two neurons. CXR: chest X-rays

from the corresponding authors if there is a reasonable request and approval from the institutional review boards of the affiliated institutions.

Performance evaluation and statistical analysis

Differences in demographic, clinical, and laboratory variables between the training and testing sets and between patients who had died and who had survived were assessed using student's *t*-test for continuous variables and chi-square test for categorical variables. Results are presented as median (interquartile range [IQR]) for continuous variables and as number (percentage) for categorical data. A two-sided $p < 0.05$ was considered statistically significant. Model performance was evaluated with area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, and F1-score.

Role of the funding source

The funding source had no role in the study design, data collection, data analysis, interpretation, or writing of the report. All the authors have full access to the data and take full responsibility for the contents of this report and the decision to submit it for publication.

Results

Subjects and clinical outcomes

Retrospective chart review identified 546 patients at University of Pennsylvania–affiliated hospitals and 108 at Brown University–affiliated hospitals meeting inclusion

criteria. Patients from the University of Pennsylvania were designated as the training set, and patients from Brown University were designated as the testing set. A total of 5645 CXRs were available for analysis (Figure 2). The median number of CXRs per patient was 4 (IQR 2–10) in the training set and 6 (IQR 2–14.5) in the testing set. The median number of days from the last ICU CXR to death was 2 (IQR 1–5) for the training set, 2 (IQR 1–6) for the testing set, and 2 (IQR 1–5) for all patients. There were no statistically significant demographic or clinical differences between the training and testing sets, except for a higher incidence of chronic kidney disease in patients in the training set ($p = 0.016$). Of the patients included in both the training and testing sets, 212 (32%) had died and 442 (68%) had survived. A higher proportion of patients died in the testing set as compared to the training set ($p = 0.010$). In terms of laboratory variables at the time of ICU admission, a larger proportion of deceased patients had elevated absolute WBC count ($p = 0.0092$), decreased absolute lymphocyte count ($p = 0.0054$), and elevated creatinine concentration ($p < 0.001$). In terms of comorbidities, deceased patients had a higher incidence of CVD ($p = 0.015$), COPD ($p = 0.0019$), and chronic kidney disease ($p = 0.036$). A detailed summary of demographic, laboratory, and clinical variables is provided in Tables 1 and 2.

Performance results

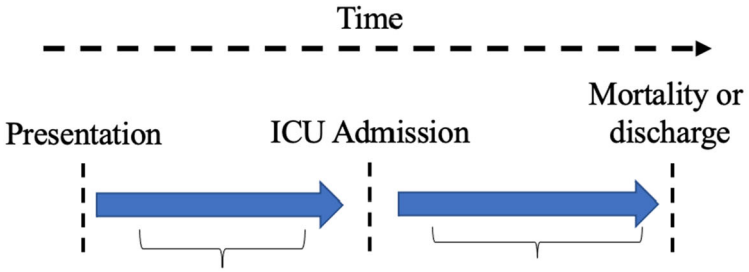
The model based on clinical data only achieved an AUC of 0.653 and an accuracy of 0.657. The model based on pre-ICU CXRs achieved an AUC of 0.632 and an accuracy of 0.593, while the model based on ICU CXRs achieved an AUC of 0.697 and an accuracy of 0.657. The longitudinal model, which considered both pre-ICU and ICU CXRs, achieved an AUC of 0.702 and an accuracy of 0.694. The addition of longitudinal CXRs significantly improved performance compared to the clinical-only model ($p = 0.039$), as the combined model reached an AUC of 0.727, an accuracy of 0.732, a sensitivity of 0.714, a specificity of 0.746, and an F1-score of 0.707. A detailed summary of model performance is shown in Table 3, Table 4, and Figure 3.

Prognostic values of clinical variables

The relative importance of clinical features was investigated in the training (Figure 4) and testing (Figure 5) datasets using RF models. In both datasets, age was found to be highly prognostic of mortality risk, followed by the presence or absence of comorbid CVD and elevated creatinine concentration. Of the other comorbidities considered, the least important were HIV and chronic liver disease. Comorbid COPD was found to be only moderately important and was more important in the training than the testing dataset.

Discussion

This study demonstrates that a deep learning model based on longitudinal CXRs and clinical information performs well in predicting in-hospital mortality of COVID-19 patients in the ICU. Five separate LTBN models were trained, and their performances were compared. The longitudinal imaging model, which included all CXRs from the time of ED presentation to the time of death or discharge, performed slightly better than the models based on pre-ICU CXRs, ICU CXRs, and clinical



Dataset	Pre-ICU	ICU	Total
Training, N (%)	953 (22)	3467 (78)	4420
Median per patient (IQR)	1 (1-2)	3 (1-8)	4 (2-10)
Testing, N (%)	172 (14)	1053 (86)	1225
Median per patient (IQR)	1 (1-2)	5 (2-14)	6 (2-14.5)

Fig. 2 Chest X-rays in the training and testing sets. Chest X-rays were collected from the time of initial presentation to the emergency department up until either death in the ICU or discharge from the ICU. The total number of chest X-rays for each dataset that was collected before admission to the ICU (pre-ICU) and during the ICU stay are shown along with the median number per patient. N: number; IQR: interquartile range; ICU: intensive care unit

Table 1 Comparison of patient characteristics across training and test sets. All continuous variables are reported as median (interquartile range), and all categorical variables are reported as number (%). Statistically significant *p*-values are bolded ($p < 0.05$). *SpO2* oxygen saturation on room air; *WBC* absolute white blood cell count; *CVD* cardiovascular disease; *HTN* hypertension; *COPD* chronic obstructive pulmonary disease; *HIV* human immunodeficiency virus

	Training set ($n = 546$)	Testing set ($n = 108$)	*p*-value
Age (years)	66 (20)	66 (14)	0.28
Male	297 (54)	62 (57)	0.14
Dead	163 (30)	49 (45)	**0.010**
Elevated temperature (> 37 C)	369 (68)	73 (68)	0.91
Low SpO2 (< 94%)	223 (41)	52 (48)	0.19
Elevated WBC count ($> 11\times10^9$/L)	159 (29)	39 (36)	0.18
Decreased lymphocyte count ($< 1\times10^9$/L)	339 (62)	79 (73)	0.063
Elevated creatinine (≥ 1.27 mg/dL)	282 (52)	53 (49)	0.70
Comorbidities			
CVD	220 (40)	34 (31)	0.11
HTN	337 (62)	66 (61)	0.99
COPD	78 (14)	16 (15)	0.99
Diabetes	224 (41)	47 (44)	0.71
Chronic liver disease	27 (5)	3 (3)	0.46
Chronic kidney disease	134 (25)	14 (13)	**0.016**
Cancer	72 (13)	9 (8)	0.22
HIV	14 (3)	0	0.19

data only. A combined model based on longitudinal imaging and clinical data significantly outperformed one based on clinical data alone.

The proposed deep learning model has the potential to improve the triage of critically ill COVID-19 patients and improve resource allocation in the ICU. By stratifying patients by high and low risk, the model could help identify which patients should be prioritized for CT or escalation of care, particularly in resource-limited settings. Chest X-rays have advantages over CT scans, particularly for ICU patients. First, CXRs are often portable and can be performed at the patient bedside, negating the need for transportation, which could prove particularly difficult for patients requiring mechanical ventilation. Second, there is less contamination risk with CXRs. The American College of Radiology recommends a thorough cleaning of CT machines by someone wearing full

Table 2 Comparison of patient characteristics who were deceased and alive. All continuous variables are reported as median (interquartile range), and all categorical variables are reported as number (%). Statistically significant *p*-values are bolded ($p < 0.05$). *SpO2* oxygen saturation on room air; *WBC* absolute white blood cell count; *CVD* cardiovascular disease; *HTN* hypertension; *COPD* chronic obstructive pulmonary disease; *HIV* human immunodeficiency virus

	Dead ($n = 212$)	Alive ($n = 442$)	*p*-value
Age (years)	71 (17)	63 (20)	**< 0.0001**
Male	114 (54)	245 (55)	0.75
Elevated temperature (> 37 C)	140 (66)	302 (68)	0.62
Low SpO2 (< 94%)	101 (48)	174 (39)	0.054
Elevated WBC count ($> 11\times10^9$/L)	79 (37)	119 (27)	**0.0092**
Decreased lymphocyte count ($< 1\times10^9$/L)	152 (72)	266 (60)	**0.0054**
Elevated creatinine (≥ 1.27 mg/dL)	136 (64)	199 (45)	**< 0.0001**
Comorbidities			
CVD	97 (46)	157 (36)	**0.015**
HTN	131 (62)	272 (62)	0.98
COPD	44 (21)	50 (11)	**0.0019**
Diabetes	80 (38)	191 (43)	0.21
Chronic liver disease	9 (4)	21 (5)	0.93
Chronic kidney disease	59 (28)	89 (20)	**0.036**
Cancer	29 (14)	52 (12)	0.57
HIV	4 (2)	10 (2)	0.98

Table 3 Performance of mortality prediction models on the external testing set. The highest values for each metric are bolded. *AUC* area under the receiver operating characteristic curve; *CI* confidence interval. *ICU* intensive care unit

Method	AUC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	F1-score (95% CI)
Clinical model	0.653 (0.563–0.738)	0.657 (0.583–0.732)	0.592 (0.480–0.711)	0.712 (0.611–0.807)	0.609 (0.512–0.699)
Pre-ICU model	0.632 (0.539–0.713)	0.593 (0.519–0.667)	0.593 (0.479–0.707)	0.591 (0.491–0.704)	0.569 (0.469–0.655)
ICU model	0.697 (0.615–0.776)	0.657 (0.583–0.732)	0.674 (0.565–0.780)	0.644 (0.546–0.746)	0.638 (0.547–0.729)
Longitudinal model	0.702 (0.613–0.782)	0.694 (0.611–0.759)	**0.756** (**0.644–0.857**)	0.642 (0.531–0.745)	0.690 (0.593–0.771)
Combined model	**0.727** (**0.645–0.809**)	**0.732** (**0.667–0.806**)	0.714 (0.609–0.822)	**0.746** (**0.648–0.833**)	**0.707** (**0.620–0.786**)

protective equipment following each scan [23]. Moreover, CT rooms may need to be unavailable for approximately 1 h following imaging of infected patients to allow for proper air circulation [23]. Given the prevalence of COVID-19 and the ongoing burden placed on hospital systems, such a delay could lead to substantial problems with patient care. Unlike CT machines, the surfaces of portable CXRs can be easily cleaned and even transported to ambulatory care facilities when deemed medically appropriate.

This study is novel in that it considers longitudinal CXRs rather than single-timepoint imaging. The results indicate that the addition of more time-series information slightly improves model performance, as the full longitudinal model was more accurate and sensitive and had a higher AUC and F1-score than both the pre-ICU and ICU models. Several previous studies have used single-timepoint imaging acquired at the time of hospital admission to predict in-hospital mortality or disease progression with machine learning and statistical modeling approaches. For instance, a previous study by our group found that deep learning based on the initial CXR and clinical variables at presentation to the ED can predict disease severity and progression with an AUC of 0.846 and 0.792, respectively, on external datasets [16]. In another study by our group, deep learning models predicted progression to critical illness with a concordance index of 0.80 in ED patients with baseline CT and clinical data [17]. Likewise, Fang et al [24] used chest CT features to develop a severity score at baseline, which was used to train three machine learning models to predict the risk of in-hospital mortality and ICU admission. The model achieved the best AUC of 0.813 in predicting ICU admission and an AUC of 0.741 in predicting mortality. Maroldi and colleagues [25] used a semi-quantitative approach to manually score baseline CXRs at a hospital presentation. Multivariate logistic regression found that these scores correlated well with subsequent in-hospital mortality. Beyond mortality risk prediction, researchers have also used baseline imaging to predict the length of hospital stay. Wang et al [26] used a deep learning model to stratify patients by high- or low-risk groups based on hospital stay duration using baseline CT features. However, it is difficult to compare results from these studies to the present, as the present study focused specifically on ICU patients, while the previous studies included all patients admitted to the hospital.

Beyond the use of longitudinal imaging, another strength of this study is that it combines imaging and clinical variables into a single model. The results indicate that the addition of longitudinal imaging significantly improves the clinical-only model, increasing the AUC from 0.653 to 0.727 ($p = 0.039$) and the accuracy from 0.657 to 0.732. In the study by Jiao and colleagues [16], the addition of single-timepoint CXR data to the clinical-only model improved both progression and severity predictions. Still, most previous studies that have used machine learning or statistical modeling to predict in-hospital mortality have relied solely on clinical or laboratory variables rather than exploring the combination of imaging and clinical data.

Table 4 Pairwise comparison of model performance by receiver operating characteristic curves. For each comparison, the *p*-value is shown, and statistically significant values ($p<0.05$) are bolded. *ICU* intensive care unit

Method	Pre-ICU model	ICU model	Longitudinal model	Combined model
Clinical model	0.77	0.57	0.50	**0.039**
Pre-ICU model		0.33	0.21	0.13
ICU model			0.90	0.66
Longitudinal model				0.71

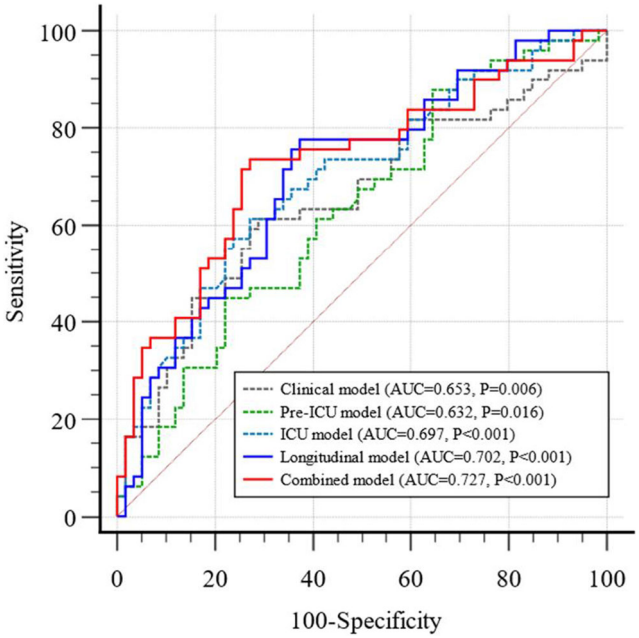


Fig. 3 Receiver operating characteristic (ROC) curves of mortality prediction models. *p*-values represent the difference from chance (AUC=0.5). AUC: area under the curve. ICU: intensive care unit

Our clinical-only model achieved a moderate performance, with an AUC of 0.653. Other models using clinical or laboratory variables to predict prognosis in COVID-19 patients have achieved better performance. For instance, Zhu et al [27] considered 78 clinical variables collected at the time of hospital presentation to predict mortality, with the top five most important variables allowing the model to achieve the best AUC of 0.968. Likewise, Hu et al [28] found that four clinical variables could predict in-hospital mortality with good accuracy, and Ko et al [29] found that an ensemble model based on deep neural networks and RFs could predict in-hospital mortality based on 28 blood biomarkers with 100% sensitivity. Another ensemble model with four machine learning methods based on 14 clinical variables was able to stratify patients by mortality risk with the best AUC of 0.976 [30]. In a large cohort, Vaid and colleagues [31] used clinical variables at admission to predict in-hospital mortality and clinical events at three, five, seven, and 10 days from admission, achieving the best AUC of 0.88 at 3 days. Multiple other studies have used similar methods [32–38].

While our clinical-only model performed worse than many in the literature, several factors should be considered. First, we only used variables that are routinely collected in the ICU to maximize the potential for integration of the model into the existing clinical workflow. Contrastively, Zhu et al [27] considered 78 total variables, and Ko et al [29] used 28 blood biomarkers. Both studies additionally used D-dimer concentrations, which we did not. A further consideration is that we chose to focus only on critically ill ICU patients. It may be inherently more difficult to predict outcomes in these patients, given that treatment recommendations evolved over the course of the outbreak. As such, patients diagnosed early may have been treated very differently from those diagnosed later, and consequently, their outcomes may be different despite similarities in baseline clinical and laboratory findings.

To identify the clinical and demographic variables that were most predictive of mortality, we performed a RF analysis. In both the training and testing datasets, age was found to be highly prognostic of mortality risk, followed by the presence or absence of comorbid CVD and elevated creatinine concentration. In other studies that performed similar analyses, age was also the most valuable factor in mortality prediction [22, 31, 33–39]. Similarly, the high concentration of CRP and the presence of one or more comorbidities were also found to be highly predictive [21, 22, 30, 31, 33, 34, 36–39]. Finally, high respiratory rate and SpO2 at admission were important for predictive accuracy [30, 33, 35, 38]. Our

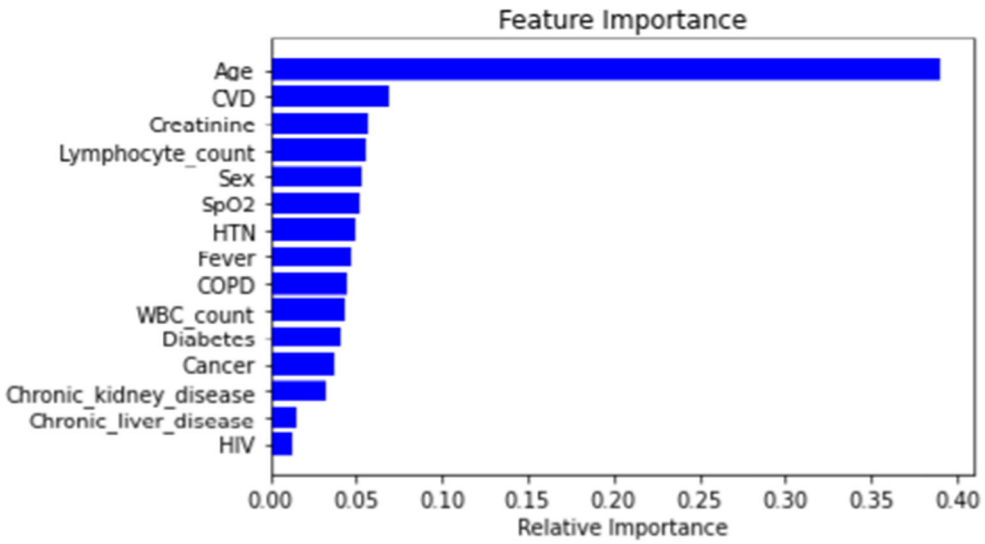


Fig. 4 Relative feature importance of clinical variables in the training data. CVD: cardiovascular disease; SpO2: oxygen saturation on room air; HTN: hypertension; COPD: chronic obstructive pulmonary disease; WBC: absolute white blood cell count; HIV: human immunodeficiency virus

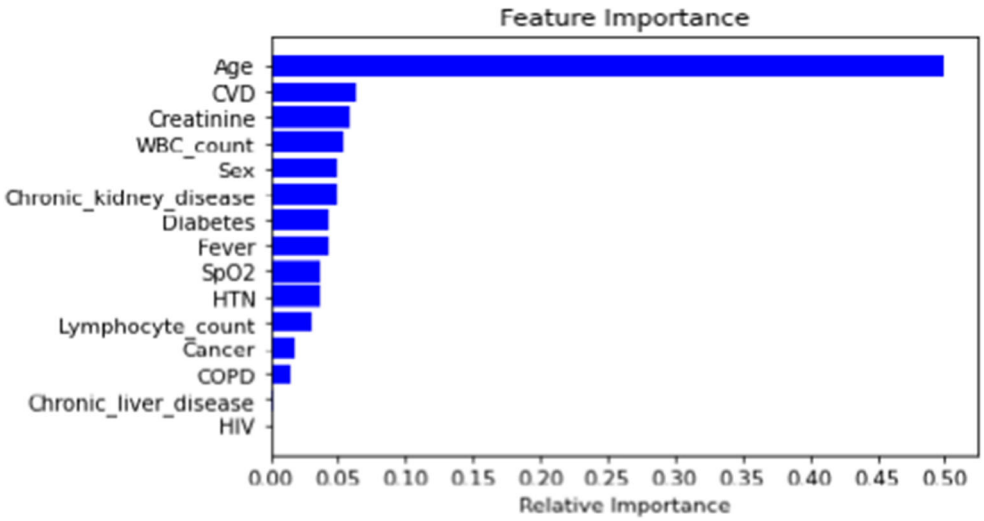


Fig. 5 Relative feature importance of clinical variables in the testing data. CVD: cardiovascular disease; SpO2: oxygen saturation on room air; HTN: hypertension; COPD: chronic obstructive pulmonary disease; WBC: absolute white blood cell count; HIV: human immunodeficiency virus

results indicate that SpO2 is only moderately important for mortality prediction compared to the other variables. Because our cohort was limited to critically ill ICU patients, a large percentage of both the survivor (39%) and non-survivor cohorts (48%) had low SpO2 (SpO2 < 94, p = 0.055). In comparison, for general COVID-19 positive patients, the average SpO2 of non-survivors is typically statistically lower (e.g., SpO2 87%) than that of survivors (e.g., SpO2 97%) [27]. Since the SpO2 was universally decreased and less variable in our cohort of ICU patients than in other studies, SpO2 was not as prognostically important in our model.

This study has several limitations. Like most machine learning models, there is a concern for generalizability, especially given that the model was trained using data obtained from a single institution, and the clinical landscape of the pandemic is quickly evolving. According to recent studies, while ICU admissions may be increasing due to increased virulence of the delta variant, death rates of ICU patients are relatively low, and there is an increasing incidence in children [4, 5, 39]. It also remains unclear whether the novel omicron variant causes more severe disease compared to infections with other variants. Moreover, the cohort in this study was entirely unvaccinated, as data collection was terminated in December 2020, and vaccines were not widely available until January 2021 [40]. It is thus unknown whether the current model will work in a highly vaccinated population with different variants of the disease. In fact, a recent study demonstrated that there was a deficit in all-cause mortality in a highly vaccinated population during the initial delta variant period from June 2021 to August 28, 2021 [41]. Given the lack of data on chest imaging characteristics of vaccinated individuals with novel variants, we are unable to predict the utility of this model on the current ICU population.

Another limitation is that we did not include treatment as a clinical variable, which may be an important consideration, given that critically ill patients are often treated aggressively and with a wide variation of approaches. Also, given the retrospective nature of data collection, there were variable numbers of CXRs collected at different times for different patients, prohibiting the development of a truly longitudinal model. Another limitation is that the median time from the last ICU CXR to death was 2 days, which provides a short window for aggressive intervention for patients identified as high risk. As shown in supplementary Figure R2, the visual appearance of CXRs from patients with different clinical outcomes was highly variable. In some cases, a radiologist would likely be able to make an accurate prediction of mortality risk with purely visual observation. Nevertheless, timely interpretation of images by a radiologist is often logistically impossible for rapidly deteriorating ICU patients during COVID-19 surges. Reading delays could potentially lead to the postponement of life-saving interventions for critically ill patients with imminent mortality risk. The automated AI model could triage incoming CXRs rapidly, allowing radiologists to prioritize workflow. Still, it would be of interest to test model performance when only early CXRs are considered (e.g., > 7 days from mortality).

While the best-performing model achieved good performance, whether the model is useful for triage in a real clinical setting is unclear. This will require prospective testing, which is a future aim. In the current model, the 95% confidence intervals for sensitivity and specificity ranged from 0.609 to 0.822 and 0.648 to 0.833, respectively. As such, at best, 17.8% of patients that progress to mortality may be missed (false negatives), and 16.7% of patients that survive may be identified as high risk (false positives). At worst, 39.1% of patients that progress to mortality may be missed (false negatives), and 35.2% of patients that survive may be identified as high risk (false positives). Despite the potential for false negatives and positives, we hypothesize that the model would serve clinically useful for rapidly triaging patients, particularly in overburdened ICUs. Finally, the models were designed to

predict mortality as a binary outcome rather than predicting overall survival time. A model which predicts not only the occurrence of mortality but also time to mortality would be of greater clinical utility.

In summary, we demonstrate that a deep learning model based on longitudinal CXRs and routinely collected clinical variables performs well in predicting in-hospital mortality of COVID-19 patients in the ICU. The addition of longitudinal CXRs improves the performance of models based on clinical data alone. Although prospective validation is required, the model has the potential to improve clinical decision-making and resource allocation for critically ill COVID-19 patients.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-022-08588-8.

Funding Research reported in this publication was partially supported by a training grant from the National Institute of Health (NIH), National Heart, Lung, and Blood Institute (NHLBI) (5T35HL094308-12, John Sollee). This research did not receive any other specific grant from funding agencies in the public, commercial, or not-for-profit sectors. All authors confirm that they have full access to all the data in the study and accept responsibility to submit the report for publication.

Data availability The data are not available for public access because of patient privacy concerns but are available from the corresponding authors if there is a reasonable request and approval from the institutional review boards of the affiliated institutions. The codebase used in this study is available online (https://github.com/chengjianhong/Covid-19-CXR.git). All implementation details are described thoroughly in the Methods and Appendix sections.

Declarations

Guarantor The scientific guarantor of this publication is Harrison X. Bai, MD

Conflict of Interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and Biometry No complex statistical methods were necessary for this paper.

Informed Consent Written informed consent was waived by the Institutional Review Board.

Ethical Approval Institutional Review Board approval was obtained.

Methodology
- Retrospective
- Diagnostic or prognostic study
- Multicenter study

References

1. Zhu N, Zhang D, Wang W et al (2020) A novel coronavirus from patients with pneumonia in China. N Engl J Med 382:727–733
2. Johns Hopkins University (2021) COVID-19 Map - Johns Hopkins Coronavirus Resource Center. Johns Hopkins University, Baltimore, MD, USA. Available via https://coronavirus.jhu.edu/map.html. Accessed 13 Dec 2021
3. Huang C, Wang Y, Li X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395:497–506
4. Bernal JL, Andrews N, Gower C et al (2021) Effectiveness of COVID-19 vaccines against the B.1.617.2 (Delta) variant. N Engl J Med 385:585–594
5. Torjesen I (2021) COVID-19: Delta variant is now UK's most dominant strain and spreading through schools. BMJ. https://doi.org/10.1136/bmj.n1445
6. Li Y, Xia L (2020) Coronavirus disease 2019 (COVID-19): role of chest CT in diagnosis and management. AJR Am J Roentgenol 214:1280–1286
7. Bernheim A, Mei X, Huang M et al (2021) Chest CT findings in coronavirus disease 2019 (COVID-19): relationship to duration of infection. Radiology 295:685–691
8. Borghesi A, Maroldi R (2020) COVID-19 outbreak in Italy: experimental chest X-ray scoring system for quantifying and monitoring disease progression. Radiol Med 125:509–513
9. Lomoro P, Verde F, Zerboni F et al (2020) COVID-19 pneumonia manifestations at the admission on chest ultrasound, radiographs, and CT: single-center study and comprehensive radiologic literature review. Eur J Radiol Open 7:100231
10. Wong HYF, Lam HYS, Fong AH-T et al (2020) Frequency and distribution of chest radiographic findings in patients positive for COVID-19. Radiology 296:E72–E78
11. Cohen JP, Dan L, Roth K et al (2020) Predicting COVID-19 pneumonia severity on chest X-ray with deep learning. Cureus 12:e9448
12. Yang W, Sirajuddin A, Zhang X et al (2020) The role of imaging in 2019 novel coronavirus pneumonia (COVID-19). Eur Radiol 30:4874–4882
13. Bai X, Wang R, Xiong Z et al (2020) Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. Radiology 296:E156–E165
14. Xu Q, Zhan X, Zhou Z et al (2021) AI-based analysis of CT images for rapid triage of COVID-19 patients. NPJ Digit Med 4:1–11
15. Borkowski A, Viswanadhan NA, Thomas LB, Guzman RD, Deland LA, Mastorides SM (2020) Using artificial intelligence for COVID-19 chest X-ray diagnosis. Fed Pract 37:398–404
16. Jiao Z, Choi JW, Halsey K et al (2021) Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study. Lancet Digit Heal 3:e286–e294
17. Wang R, Jiao Z, Yang L et al (2021) Artificial intelligence for prediction of COVID-19 progression using CT imaging and clinical data. Eur Radiol 1:205–212
18. Bai HX, Hsieh B, Xiong Z et al (2020) Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. Radiology 296:E46–E54
19. Bai HX, Wang R, Xiong Z et al (2020) Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. Radiology 296:E156–E165
20. Rigatti S (2017) Random forest. J Insur Med 47:31–39
21. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition vols 2016-December 770–778 (IEEE Computer Society, 2016). https://doi.org/10.1109/CVPR.2016.90
22. Dosovitskiy A, Beyer L, Kolesnikov A et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv. DOI: arxiv:2010.11929

23. American College of Radiology (2021) ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection. American College of Radiology, Virginia, USA. Available via https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection. Accessed 13 Dec 2021
24. Fang X, Kruger U, Homayounieh F et al (2021) Association of AI quantified COVID-19 chest CT and patient outcome. Int J Comput Assist Radiol Surg 16:435–445
25. Maroldi R, Rondi P, Agazzi GM, Ravanelli M, Borghesi A, Farina D (2020) Which role for chest x-ray score in predicting the outcome in COVID-19 pneumonia? Eur Radiol 31:4016–4022
26. Wang S, Rondi P, Agazzi GM, Ravanelli M, Borghesi A, Farina D et al (2020) A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J 56:4016–4022
27. Zhu J, Ge P, Jiang C et al (2020) Deep-learning artificial intelligence analysis of clinical variables predicts mortality in COVID-19 patients. J Am Coll Emerg Physicians Open 1:1364–1373
28. Hu C, Liu Z, Jiang Y et al (2020) Early prediction of mortality risk among patients with severe COVID-19, using machine learning. Int J Epidemiol 49:1918–1929
29. Ko H, Chung H, Kang WS et al (2020) An artificial intelligence model to predict the mortality of COVID-19 patients at hospital admission time using routine blood samples: development and validation of an ensemble model. J Med Internet Res 22:e25442
30. Gao Y, Cai G-Y, Fang W et al (2020) Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. Nat Commun 11:1–10
31. Vaid A, Somani S, Russak A et al (2020) Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: model development and validation. J Med Internet Res 20:e24018
32. Sánchez-Montañés M, Rodríguez-Belenguer P, Serrano-López AJ, Soria-Olivas E, Alakhdar-Mohmara Y (2020) Machine learning for mortality analysis in patients with COVID-19. Int J Environ Res Public Health 17:1–20
33. Abdulaal A, Patel A, Charani E, Denny S, Mughal N, Moore L (2020) Prognostic modeling of COVID-19 using artificial intelligence in the United Kingdom: model development and validation. J Med Internet Res 22:e20259
34. Guan X, Zhang B, Fu M et al (2021) Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study. Ann Med 53:257–266
35. Ikemura K, Bellin E, Yagi Y et al (2021) Using automated machine learning to predict the mortality of patients with COVID-19: prediction model development study. J Med Internet Res 23:e23458
36. Ma X, Ng M, Xu S et al (2020) Development and validation of prognosis model of mortality risk in patients with COVID-19. Epidemiol Infect 148:e168
37. Pourhomayoun M, Shakibi M (2021) Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart Heal 20:100178
38. Booth AL, Abels E, McCaffrey P (2021) Development of a prognostic model for mortality in COVID-19 infection using machine learning. Mod Pathol 34:522–531
39. Mushtaq J, Pennella R, Lavalle S et al (2021) Initial chest radiographs and artificial intelligence (AI) predict clinical outcomes in COVID-19 patients: analysis of 697 Italian patients. Eur Radiol 31:1770–1779
40. American Journal of Managad Care (2021) A timeline of COVID-19 vaccine developments in 2021. The American Journal of Managed Care, Cranbury, NJ, USA. Available via https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-developments-in-2021. Accessed 13 Dec 2021
41. Faust JS, Du C, Maye KD et al (2021) Absence of excess mortality in a highly vaccinated population during the initial COVID-19 Delta period. medRxiv. https://doi.org/10.1101/2021.09.16.21263477