Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Identifying Risk of Adverse Outcomes in COVID-19 Patients via Artificial Intelligence-Powered Analysis of 12-Lead Intake Electrocardiogram

Arun R. Sridhar¹*, Zih-Hua Chen², Jacob J Mayfield¹, Alison E Fohner³, Panagiotis Arvanitis⁴, Sarah Atkinson¹, Frieder Braunschweig⁵,⁶, Neal A Chatterjee¹, Alessio Falasca Zamponi⁴, Gregory Johnson⁷, Sanika A Joshi², Mats C H Lassen⁸, Jeanne E Poole¹, Christopher Rumer¹, Kristoffer G Skaarup⁸, Tor Biering-Sørensen⁸*, Carina Blomstrom-Lundqvist⁴*, Cecilia M Linde⁵,⁶*, Mary M Malekar⁹*, and Patrick M. Boyle²,¹⁰,¹¹*

¹. Division of Cardiology, University of Washington, Seattle, WA, USA.
². Department of Bioengineering, University of Washington, Seattle, WA, USA.
³. Department of Epidemiology, University of Washington, Seattle, WA, USA.
⁴. Department of Medical Science, Uppsala University, Sweden.
⁵. Karolinska Institutet, Department of Medicine, Stockholm, Sweden.
⁶. Karolinska University Hospital, Department of Cardiology, Stockholm, Sweden.
⁷. Unaffiliated independent researcher, Seattle, WA, USA.
⁸. Department of Cardiology, Herlev & Gentofte University Hospital, Copenhagen University, Denmark.
⁹. Department of Computational Physiology, Simula Research Laboratory, Oslo, Norway.
¹⁰. Institute for Stem Cell and Regenerative Medicine, University of Washington, Seattle, WA, USA.
¹¹. Center for Cardiovascular Biology, University of Washington, Seattle, WA, USA.

*Corresponding authors: arun11@cardiology.washington.edu AND pmboyle@uw.edu

Site principal investigators

Conflicts of interest: ARS and PMB report that during the initial data collection phase at UW, GE provided access to the MUSE plugin necessary for exporting ECG data in machine-readable format free of charge; otherwise, GE was not involved in this project. CL reports consulting fees from AstraZeneca, Roche diagnostics, Bayer and speaker honoraria from Novartis, Astra, Bayer, Medtronic, Impulse Dynamics and Vifor and receives funding from Swedish Heart-Lung-foundation, from Stockholm County Council and from the Royal Society of Science. TBS reports receiving research grants from Sanofi Pasteur, and GE Healthcare, is a Steering Committee member of the Amgen financed GALACTIC-HF trial, on advisory boards for Sanofi Pasteur and Amgen, and speaker honorariums from Novartis and Sanofi Pasteur. The remaining authors have nothing to disclose.

Acknowledgments: This work was supported by a COVID-19 Rapid Response Grant from the University of Washington Population Health Initiative (awarded to ARS, AEF, NAC, JEP, and PMB). The work was further supported by Afa Insurance in Sweden, Dnr 200223, Novo Nordisk Foundation in Denmark (NFF20SA0062835), and Norwegian Center for Research-Based Innovation ProCardio SFF. The REDCap database system, which was essential to the completion of this project, was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1 TR002319. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors are grateful to research nurse Helena Karlsson (KI) for assistance with data entry for this project.
Abstract

**Background:** Adverse events in COVID-19 are difficult to predict. Risk stratification is encumbered by the need to protect healthcare workers. We hypothesize that AI can help identify subtle signs of myocardial involvement in the 12-lead electrocardiogram (ECG), which could help predict complications. **Objective:** Use intake ECGs from COVID-19 patients to train AI models to predict risk of mortality or major adverse cardiovascular events (MACE). **Methods:** We studied intake ECGs from 1448 COVID-19 patients (60.5% male, 63.4±16.9 years). Records were labeled by mortality (death vs. discharge) or MACE (no events vs. arrhythmic, heart failure [HF], or thromboembolic [TE] events), then used to train AI models; these were compared to conventional regression models developed using demographic and comorbidity data. **Results:** 245 (17.7%) patients died (67.3% male, 74.5±14.4 years); 352 (24.4%) experienced at least one MACE (119 arrhythmic; 107 HF; 130 TE). AI models predicted mortality and MACE with area under the curve (AUC) values of 0.60±0.05 and 0.55±0.07, respectively; these were comparable to AUC values for conventional models (0.73±0.07 and 0.65±0.10). There were no prominent temporal trends in mortality rate or MACE incidence in our cohort; holdout testing with data from after a cutoff date (June 9, 2020) did not degrade model performance. **Conclusion:** Using intake ECGs alone, our AI models had limited ability to predict hospitalized COVID-19 patients’ risk of mortality or MACE. Our models’ accuracy was comparable to that of conventional models built using more in-depth information, but translation to clinical use would require higher sensitivity and positive predictive value. In the future, we hope that mixed-input AI models utilizing both ECG and clinical data may be developed to enhance predictive accuracy.

**Keywords:**

12-lead ECG • COVID-19 • Artificial Intelligence • Deep Learning
Arrhythmia • Heart Failure Prognosis • Mortality • Risk Factors
Introduction

Coronavirus disease of 2019 (COVID-19) has now been documented in at least 230 million people worldwide, resulting in at least 4.7 million deaths.\(^1\)-\(^6\) This affliction has been linked with a range of cardiovascular complications including arrhythmias, myocarditis, acute coronary syndrome (ACS), and thromboembolism.\(^7\) These \textit{de novo} cardiovascular events, as well as pre-existing cardiovascular comorbidity are linked to adverse outcomes in patients with COVID-19. Work in cardiomyocytes derived from human pluripotent stem cells showed that SARS-CoV-2 infection impaired electrophysiological and contractile function and led to widespread cell death, suggesting cardiovascular symptoms may be a direct consequence of cardiotoxicity.\(^8\) Atrial fibrillation (AF)\(^9\) is associated with increased mortality in infected patients, especially when it newly presents during hospitalization for COVID-19.\(^5\) Biomarker-detected myocardial injury,\(^10\) thromboembolism, and abnormal laboratory coagulopathy studies are frequent complications of SARS-CoV-2 infection and are also associated with worse outcomes.\(^11\)-\(^13\) Early prediction of mortality and morbidity risk may improve patient management.

One of the greatest challenges during the pandemic, especially in resource-strained settings, has been early identification of individual patients at higher risk for adverse outcomes. Machine learning (ML) has the potential to facilitate triage by providing clinicians with useful risk stratification data to inform decisions regarding level of care and follow-up monitoring. The electrocardiogram (ECG), a rapid, inexpensive, and non-invasive diagnostic test suitable for repetitive recordings, is an ideal target for ML augmentation, especially in light of associations observed between ECG abnormalities and adverse COVID-19 outcomes.\(^14\),\(^15\) ECG ML applications have been shown to recognize subtle patterns in the electrical signals, imperceptible to human readers, that can be leveraged to predict and classify different arrhythmic conditions, including AF,\(^16\) and to screen for other cardiovascular conditions, including hypertrophic cardiomyopathy,\(^17\) left ventricular systolic dysfunction,\(^18\) and aortic stenosis.\(^19\) ML networks have also demonstrated promise identifying clinically meaningful markers of prognosis, predicting both 1-year mortality\(^20\) and incident cardiac arrest within 24 hours.\(^21\)
In this study, we developed a deep neural network (DNN)\textsuperscript{22,23} based ECG analysis system to predict the risk of adverse events in patients with COVID-19 solely using 12-lead intake ECGs recorded at the time of COVID-19 hospital admissions as input. We trained two DNNs, one to predict mortality and one to predict major adverse cardiovascular events (MACE, including arrhythmic events, new onset heart failure, or thromboembolic complications). We report the performance of these DNN systems, which use the ECGs alone to predict these outcomes of interest and compare these to a conventional statistical model which uses a broad set of demographics, comorbidities, and clinical variables to make the same predictions.

Methods

Collaboration Setup

We enrolled patients from the University of Washington Healthcare System (UW; Seattle, WA), the Karolinska Institutet University Hospital (KI; Stockholm, Sweden), the Uppsala University Hospital (UU; Uppsala, Sweden), and the Copenhagen University Hospital (UC; Copenhagen, Denmark). Data collection was retrospective for all centers except UC, where collection was prospective, but data were analyzed retrospectively. This study was approved by the Institutional Review Board (IRB) of the UW, the Swedish Ethical Review Authority, and the regional Danish Ethical Committee. Associated reference numbers: IRB6878 (UW), 2020-02627 (KI), 2020-02662 (UU), and H-20021500 (UC).

Dataset Design

We first identified 1448 patients who were admitted to one of the four study centers between March 2, 2020, and February 28, 2021 for COVID-19; those who underwent an ECG recording during hospital intake were included. All included patients had laboratory-confirmed COVID-19 diagnosis (ICD-10-CM U07.1). Each center identified records from March 2, 2020 to June 8, 2020. UW, which was the main center for the study, continued extracting records through Feb 28, 2021. ECGs were adjudicated to exclude cases with pacing artifacts caused by cardiac implantable electronic devices (CIED) or unacceptable quality (due to missing leads or significant noise); in total, 62 records were excluded due to these criteria (24 pacing artifacts; 38 poor quality). The remaining 1386 hospitalized COVID-19 patients
(60.5% male, 63.4±16.9 years old) comprised our final dataset. Standard 10-second 12-lead ECGs and inpatient outcomes (see next subsection) were collected for the study cohort, along with demographics, comorbidities, hospitalization variables, and laboratory values (see Table 1 for details of variables collected). ECGs were acquired using GE ECG systems (General Electric Company, USA) and raw data were managed using the GE MUSE Cardiology Information System. Twelve hundred and seven of 1386 (95.0%) intake ECGs were acquired within 24h of the hospital admission date; for the remaining 179 patients, the median number of days between hospital admission and intake ECG acquisition was 4 (inter-quartile range: 3, 8). For patients who had multiple ECGs in the MUSE system from date of intake, the latest one was used in the dataset. All data were collected and managed using Research Electronic Data Capture (REDCap) electronic data capture tools hosted at the UW Institute of Translational Health Sciences. Data entry was overseen or performed by experienced research nurses or doctors at all four sites. A PDF copy of the REDCap Data Collection Instrument used to populate our database is provided in the supplemental material.

**Clinical Outcomes**

Records were labeled by clinician and/or trained study coordinators with clinician supervision for (1) patient all-cause mortality vs. survival and (2) incidence of major adverse cardiovascular events (MACE) during COVID-19 hospitalization (controls without events vs. cases with arrhythmic, heart failure [HF], or thromboembolic events). Co-primary endpoints were mortality and composite of MACE which included thromboembolic, arrhythmic, or HF events. Thromboembolic events were defined as acute myocardial infarction, ischemic stroke, or pulmonary embolism. Arrhythmic events were defined as new onset AF, high burden of premature ventricular complexes (PVCs), sustained ventricular tachycardia (VT), ventricular fibrillation (VF) and cardiac arrests due to brady- and tachy-arrhythmias. HF events included new-onset HF and cardiogenic shock. Patients with previously known or recurrent AF were noted, but these were not classified as events under the arrhythmic endpoint since our aim was to predict complications that arose *de novo* because of COVID-19. The non-event cases formed the background class, in which patients developed no MACE during their COVID-19 hospitalization. The study period
only comprised the hospitalization for COVID-19 and no long term follow up was performed after discharge. Thus, all included patients either died or were discharged.

**AI Model Development**

Essential details about the AI models are provided in this section; see Supplemental Methods for further information. Figure 1 presents a dataflow diagram for our study. In addition to the general exclusion criteria described above, 6 records were excluded from the MACE network analysis due to the presence of AF on the intake ECG in a patient with new onset AF as an outcome. ECGs and associated labeling information were split into training and testing sets using a 10-fold stratified cross validation scheme. The ratio of records in the training, validation, and testing sets was 8:1:1. The deep learning model input data was a stack of standard 10-second 12 lead ECG records sampled at either 500 (n=1217) or 250 Hz (n=169). For each clinical record used the model input was an 8 x 5000 matrix.

We formulated two distinct CNN-LSTMs; these were designed to predict the incidence during COVID-19 hospitalization of, respectively, mortality (CNN-LSTM1) and life-threatening cardiovascular events as described above (CNN-LSTM2). Figure 2 shows schematic representations for the model architectures, both of which had three main sections. In section I, a series of 1D convolution layers was used to extract temporal features from each channel separately. Every convolution layer was followed by a batch normalization layer to centralize data distribution and a rectified linear unit (ReLU) to weight the output of past layers. Section II was a recurrent neural network (RNN), consisting of two LSTM layers to process spatial features across all eight channels; each LSTM layer contained up to four units with feedback connections. Section III was a fully connected layer and activation function, used to output a value describing the model’s confidence that a particular ECG belonged to each class. CNN-LSTM1 was a binary model (i.e., predicted likelihood of survival vs. death), whereas CNN-LSTM2 was a multi-label model (independent likelihoods for all four event types: arrhythmic, heart failure, thromboembolic, or none).

In addition to the primary analysis via 10-fold cross validation, a secondary analysis was conducted to evaluate whether the timing of patient intake might affect the predictive power of our AI models. In this case, we trained the models with only ECGs of patients who were admitted between March 2, 2020, and June 8, 2020. Model performance was then evaluated with a hold-out test set that
contained only ECGs of patients who were admitted on or after June 9, 2020. Lastly, to explore the possibility that changing the cohort size might affect model accuracy, we re-ran the entire 10-fold train/validate/test process for CNN-LSTM1 for subsets with 20%, 40%, 60%, or 80% of the population. These were used to construct a learning curve (i.e., model accuracy vs. cohort size), as in prior work.28

Training and testing of CNN-LSTM models was carried out using advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system of University of Washington. All jobs were run on one standard compute node (32 cores, Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz, 128 GB RAM).

Statistical Assessment of AI Model Performance

Model performance was evaluated by calculating area under the receiver operator characteristic curves (AUROC), sensitivity and specificity values, as well as confusion matrices for hold-out testing sets in a 10-fold cross validation scheme, as described above. Optimal probability thresholds were determined based on the ROC curves of internal validation sets for each output class:

\[ D = f_{pr_{val}}^2 + 2 \times (1 - tpr_{val})^2 \]

where \( D \) is a weighted distance from the origin of the coordinate system to a specific point in the ROC curve of validation set, \( fpr \) is false positive rate, and \( tpr \) is true positive rate. This formulation intentionally weighted sensitivity (i.e., \( 1 - tpr \)) over specificity (i.e., \( fpr \)), since the importance of minimizing false negative rate (\( fnr \)) was deemed a much higher priority than reducing \( fpr \). Each \( fpr, tpr \) pair maps to a unique threshold determined by minimizing \( D \).

The ROC curves for each class in CNN-LSTM2 independently consider the comparison of model output to ground truth label. Thus, assessing the overall performance of this model necessitated the use of an ROC aggregation technique. We thus calculated the macro-average ROC curve (i.e., each class’s ROC curve contributed equally to the average), which considered each prediction problem separately (e.g., survival vs. non-survival AND death vs. non-death, independently) then averaged them together. Finally, when aggregating results of different model performances from all different train/test iterations carried out as part of the 10-fold cross validation process, two-sided 95% confidence intervals were used to estimate the overall performance matrix for the system.
**Conventional Statistical Model Development**

Clinical model comparisons were developed using multivariable logistic regression. The cohort was split into the same 10 training/testing sets used for the AI models. The endpoints of mortality and composite cardiac event (experience of at least one: heart failure, arrhythmia, or thromboembolism) were modeled separately. The following steps were performed for each of the 10 cohort splits: Missing data was imputed in the training set using mean imputation and the same mean values were applied to missing data in the test and validation sets. Logistic regression was used to predict the outcome. The models adjusted for age, sex, race/ethnicity, body mass index (BMI), history of hypertension, history of coronary artery disease, and presence of an intracardiac device. Four metrics extracted from ECGs were also included: ventricular rate, PR Interval, QRS duration, and QT interval. ECG metrics were treated as continuous variables. Race categories were Asian, Black, White Hispanic, White non-Hispanic, and Other/Unknown. BMI was either entered by clinical staff or calculated from height and weight. Age and BMI were treated as continuous variables. History of hypertension, history of coronary artery disease, and presence of intracardiac device were treated as binary variables, with the reference being not present. The reference category for sex assigned at birth was male. Race/ethnicity was included as a factor, with non-Hispanic White as the reference. The logistic regression models used the same optimal threshold identification process as outlined above for the AI models in the test set and performance metrics were evaluated in the validation cohort. Results of the model performances from all different train/validation/test iterations carried out as part of the 10-fold cross validation process were aggregated to determine two-sided 95% confidence intervals. Analyses were performed using the R programming language version 3.6.1 and used packages: readxl, tidyr, dplyr, caret, and pROC.

**Results**

We used intake ECGs and outcomes during COVID-19 hospitalization for 1386 patients, after all exclusion criteria had been applied (see Methods). All four centers were University hospitals (Table 1). The mean age was 63.43±16.9 years; 60.5% (n=839) were assigned male sex at birth, 54.0% (n=737 of 1386).
had hypertension, and 1.9% (n=27 of 1386) had a CIED (22 [1.6%] pacemakers; 5 [0.4%] implantable cardioverter defibrillators). 116 intake ECGs included atrial arrhythmia, which accounted for 8.4% of the total database. Ninety-seven were AF and 12 were atrial flutter, with various other arrhythmias accounting for the remaining cases (Supplemental Table 1); of note, six records with new onset AF were excluded from the development of CNN-LSTM2 to avoid confounding, as outlined in Methods.

The all-cause mortality rate was 17.7% (n=245); the survival group included patients who were discharged either to home (n=776; 63.7%) or to other care units after stabilization (n=235; 19.3%). As shown in Table 1, sub-cohorts from different regions were distinct, with statistical tests showing a lack of homogeneity in all demographic and comorbidity categories except hypertension. In contrast, outcomes were more uniformly distributed across the centers, with exceptions being a higher incidence of HF events at KI and a lower mortality rate at UC. We also tabulated these data, along with ECG metrics, according to mortality vs. survival outcome (Supplemental Table 2) and event type (none vs. arrhythmic vs. HF vs. thromboembolic event; Supplemental Table 3). Compared to those who survived, patients who died were older (average age: 74 vs. 61 years) and had higher prevalence of hypertension (70% vs. 50%) and coronary artery disease (CAD) (23% vs. 12%); for comparison between event types, significant inter-group differences existed for several demographic variables as well as hypertension and CAD (both of which were higher in patients who experienced arrhythmia or HF events). In two sub-groups (patients who died vs. those who survived and patients with HF vs. no MACE), ventricular rate was faster, QRS duration was greater, and corrected QT interval was prolonged (P<0.01 in all cases). Given the high rate of atrial arrhythmia in this cohort, it is notable that there were no significant inter-group differences whatsoever in PR interval or P wave duration.

The ability of the fully trained CNN-LSTM1 model to successfully predict death during hospitalization vs. survival from intake ECG is illustrated by the ROC curve in Fig. 3A (green line; AUROC: 0.60±0.05); the ROC curve for the conventional statistical model is also shown for comparison (red line; AUROC: 0.73±0.0.07). Additional detail regarding the conventional model (odds ratio and P-values for each covariate in the model) are presented in Supplemental Table 4. In both cases, the line shown is the average across all 10 cross-validation data sets and the shaded region shows the ±1
standard deviation range. The predictive power of \( \text{CNN-LSTM1} \) was not approved by either of the transfer learning (TL) approaches described in the Supplemental Methods (TL1 AUROC: 0.42±0.06; TL2 AUROC: 0.54±0.06). At the optimal threshold value, positive and negative predictive (PPV/NPV) values for \( \text{CNN-LSTM1} \), with [95% confidence intervals], were 0.22 [0.20-0.23] and 0.87 [0.85-0.89]; the overall sensitivity and specificity were 0.66 [0.60-0.73] and 0.47 [0.41-0.53]. These values were distinct from those observed for the conventional statistical model, in which positive and negative predictive powers were inverted (PPV: 0.90 [0.88-0.92]; NPV: 0.35 [0.32-0.38]; sensitivity: 0.75 [0.69-0.80]; specificity: 0.61 [0.50-0.72]). To further examine the tradeoff between highly undesirable \( fnr \) (i.e., survival predicted for a patient at risk of dying) and more tolerable \( fpr \) (i.e., death predicted for a patient tracking towards survival), we plotted both quantities as a function of the raw \( \text{CNN-LSTM1} \) output threshold for distinguishing between predicted death and survival (Fig. 3B). This analysis shows that to maintain a reasonable \( fnr \) (e.g., 20% for the nominal threshold shown by the dashed yellow line), it was necessary to accept a high \( fpr \) (~70% for the example shown); this is consistent with the high NPV and low PPV values reported above. Construction of a learning curve (Supplemental Fig. 1) for \( \text{CNN-LSTM1} \) showed that there was no trend towards increasing model accuracy as a function of cohort size. Rather, there was a trend towards reduced accuracy, although difference in accuracy for successive cohort sizes were not statistically significant.

Summary data (i.e., macro-average ROC curve; AUROC: 0.55±0.07) for prediction of cardiovascular events via \( \text{CNN-LSTM2} \) are shown in Fig. 3C (green line). These results are contrasted with those from the logistic regression model shown in Fig. 3C (red line), which tackled a distinct classification problem that was binary (i.e., MACE vs. no MACE; AUROC: 0.65±0.10); see Supplemental Table 4 for additional covariate information from this model. Both the ML-based and conventional approach had inferior power compared to the respective death vs. survival models, indicating that MACE prediction (multi-class using \( \text{CNN-LSTM2} \) or binary using the conventional model) was more challenging. Quantitative metrics of the \( \text{CNN-LSTM2} \) aggregate performance were as follows: PPV = 0.28 [0.27-0.29]; NPV = 0.76 [0.74-0.78]; sensitivity = 0.67 [0.61-0.72]; specificity = 0.42 [0.36-0.47]. The predictive power of the logistic regression model was comparatively higher: PPV: 0.82 [0.81-0.84]; NPV: 0.32 [0.26-0.38];
sensitivity: 0.75 [0.69-0.82]; specificity: 0.39 [0.29-0.49]. Since TL was not beneficial in the context of the mortality vs. survival network, we elected not to attempt it for CNN-LSTM2.

ROC curves illustrating model performance for independent prediction of the four individual MACE types are provided in Fig. 4A-D. Here, we see that poor performance of CNN-LSTM2 was disproportionately driven by the model’s difficulty in identifying ECGs from patients who had no event or experienced a TE (AUROC = 0.54±0.07 and 0.51±0.06, respectively); in contrast, the model’s ability to predict arrhythmic or HF events was on par with the AI-based ability to predict death (AUROC = 0.58±0.07 and 0.59±0.07, respectively). Nevertheless, as shown by graphs of \( fpr \) and \( fnr \) as a function of event-specific model thresholds for all four outcome types in Supplemental Fig. 2A-D, the practical consequence of this increased predictive power is modest (i.e., dashed black lines showing the \( fpr \) that must be tolerated to achieve a 20% \( tpr \) all range between ~75% and ~80%, with little difference between event types).

CNN-LSTM1 outputs for predicted death and survival probabilities for intake ECGs in a representative test set are shown in Fig. 5A. As indicated by the box-and-whisker diagrams superimposed on the violin plots, the optimal threshold value (dashed line) resulted in 72% of the ECGs from patients who died being assigned a “death” prediction (i.e., \( tpr \); Fig. 5A, right). The corresponding rate of correct classification for ECGs from patients who survived was lower (Fig. 5A, left: 54% \( tnr \)). Similar plots for CNN-LSTM2 (Figs. 5B-C and Supplemental Figs. 3A-B) indicate that the multi-label classification problem was more challenging.

Finally, as shown in Fig. 6, using a test set comprising only ECGs and outcomes from patients recruited after a cutoff date (June 9, 2020), the predictive power of both the survival vs. death model (Fig. 6A) and the multi-label MACE model (Fig. 6B; macro-average and individual predictive ROCs) were within the ±1 standard deviation intervals for the corresponding models shown in Fig. 3A and 3C, respectively. The numbers of patients enrolled at each center over the study period are shown in Supplemental Figs. 4A-B. There were no discernable trends in mortality rate or MACE incidence on a month-by-month basis (see Supplemental Figs. 4C and D, respectively).
Discussion

In this manuscript, we demonstrate the feasibility and examine the potential effectiveness of a rapid, inexpensive triage tool for COVID-19 patients using a 12-lead ECG at admission augmented by machine learning. To our knowledge, this is the first such attempt to triage any viral disease using a single ECG. Destabilization of healthcare infrastructure during the pandemic sharply highlighted the need for rapid and readily available triage tools to guide resource allocation during times of crisis. When compared with the traditional statistical model incorporating extensive clinical data, our results using only the DNN-augmented ECG did not yield improved discrimination, however, this approach has the advantage of not requiring clinical expertise to gather medical history. Nevertheless, our models’ low sensitivity and positive predictive value would be a barrier to successful deployment in clinical settings in its current form. Mixed-input AI models that analyze both ECG data and clinical variables might enhance the accuracy, as discussed below.

Our network yielded AUROC values of 0.60 for predicting death and 0.55 for predicting MACE. The overall performance of our models can be compared to other commonly used risk stratification tools such as the CHADS$_2$ and CHA$_2$DS$_2$-VASc scores for stroke which are estimated by the U.S. Agency for Healthcare Research and Quality to have AUROC values of 0.66 to 0.75. Due to the unpredictable nature of the disease, especially in hospitalized patients, and the high risk of devastating complications, we optimized our DNN to maximize NPV. This prioritized correct identification of patients who had favorable outcomes as low risk (i.e., true negatives), while minimizing false negatives. Our model achieved a NPV of 0.87 for mortality and 0.76 for MACE.

Of the two networks we developed, CNN-LSTM1 had distinctly better performance than CNN-LSTM2. Although the specifics of the classification tasks involved are distinct (multi-class vs. binary), the general implication is that intake ECG data may be less suitable for predicting MACE during COVID-19 hospitalization compared to demographic data and comorbidity data. Nevertheless, results from CNN-LSTM2 must be carefully interpreted since its ability to predict different MACE types was divergent. For the best performing individual prediction (HF event classifier), at the optimal threshold the network had
relatively low $f_{pr}$ (e.g., 44% for ECGs from patients with no event; Fig. 5C, column 1) but this came at the cost of a relatively high $f_{nr}$ (27%; Fig. 5C, column 4). In the case of independently predicting the likelihood that a patient would experience no MACE (Supplemental Fig. 3A), while 73% of ECGs from patients who truly lacked adverse events were correctly classified (column 1), >50% of patients in all three MACE groups were incorrectly classified (columns 2-4).

The strategic holdout analysis presented in Fig. 6 also warrants further discussion. The aim here was to determine if the model’s predictive power was degraded by attempting to train and test the model with ECGs from patients hospitalized earlier and later in the pandemic, respectively. This might be expected, due to the changing attributes of COVID-19 disease and treatment over time, although our analysis of temporal distribution of mortality rates and MACE incidence in this cohort (Supplemental Figs. 4C-D) showed no apparent trends. In practice, holdout model performance was not obviously inferior to the main results presented in Fig. 3. Notably, the holdout MACE model’s ability to predict HF from intake ECG was among the highest observed in the study (AUROC = 0.68).

Even though our enthusiasm for the potential translational value of these particular models is low, our study has significant strengths. We initiated an international collaboration of four university hospitals, which were heavily affected by the pandemic across two continents. This yielded a diverse set of patients from different geographic regions and institutions with significant heterogeneity in presentation as well as therapies. Various conventional machine learning techniques have been proposed for ECG-based algorithms, but most use ECG features, as opposed to the raw ECG signal itself as employed here, as input (e.g., RR interval, QRS complex duration, etc.), which requires labor-intensive pre-processing. Moreover, we opted for minimal data sanitization by enforcing only the most essential exclusion criteria and applying minimal pre-processing to model inputs. Thus, the data used to train and test our ML networks closely resemble “real world” signals that could be found in any clinical setting. We highlight that atrial arrhythmia is common and likely to be encountered in many patients presenting with COVID-19, so we opted to include as many of these ECGs as possible in our study. The lone exception was that ECGs demonstrating new-onset AF were excluded from the MACE outcome, since new onset AF was a component of the MACE outcome.
It is also worth noting that CNN-LSTM2 carries out a multi-label classification task in which the likelihood of a particular ECG belonging to each group (no event vs. three types of MACE) is independently predicted via one single set of calculations. This is a distinct and more difficult classification task from the conventional statistical model against which the performance of CNN-LSTM2 was evaluated (i.e., binary prediction of no MACE vs. any MACE). Finally, we carried out a rigorous and comprehensive hyper-parameter tuning exercise for both CNN-LSTM1 and CNN-LSTM2. In each case, we considered 128 unique network permutations and carried out a complete set of train/test cycles using our 10-fold cross-validation schemes (i.e., 1280 unique runs per model), meaning that the predictive capabilities of the models shown in Fig. 2 and summarized in Supplemental Tables 4-5 are the best possible outcome for this set of DNN parameters. As such, we are confident that the modest AUROC values for our two models are not the consequence of an inadequate exploration of what DNN technology might potentially offer.

The imperfect performance of our network demonstrates an important aspect of ML applications: explainability. While many DNN architectures allow for identification of specific features predictive of adverse outcomes, we were unable to do so because of the LSTM components included in our DNNs. The CNN models in our study were trained to identify adverse outcomes regardless of the exact changes occurring on the ECG. We believe it is entirely possible that the CNNs could be using any changes encoded in that signal (e.g., heart rate, heart rate variability, QRS width, ST changes, or combinations of these elements) to make predictions about outcomes. In the case of COVID-19, it is possible that the most salient ECG markers are those that are indicative of baseline comorbidity, such as right atrial abnormality or right ventricular strain pattern which suggest pulmonary disease. Alternatively, the network may have preferentially identified novel features associated with severe COVID-19 disease. To examine exactly what features the CNN models are examining to make these predictions, we would have to use certain explainable AI models, which would be more interpretable. This was beyond the scope of the current project, but future studies may be able to harness new computational techniques to improve approaches in this regard. It is also difficult to speculate what steps may be taken to improve model performance. We hypothesize that the heterogeneity of both the components of MACE and varied causes
of mortality contributed to poor performance. Future efforts may benefit from identification of more narrowly defined outcomes with linear and biologically plausible mechanistic connections between cardiac electrical activity and the outcome.

Some interesting observations can be gleaned from the odds ratios for covariates in the two conventional models used as a basis of comparison for the AI-based models. Advanced age, male sex, and higher ventricular rate were associated with increased risk of both mortality and MACE; for the second model, there were additional associations between MACE risk and elevated BMI or increased PR interval duration, albeit with weaker $P$-values. The clinical significance of these features is not entirely clear. In the context of the current study, it is noteworthy at least one ECG feature (ventricular rate) was strongly associated with both death and MACE. While it is possible that formulate of alternative statistical models using exclusively ECG metrics, demographics data, or co-morbidity information might elucidate the key drivers of predictive capability, these analyses were deemed beyond the scope of this AI-focused study. Nevertheless, the fact that the predictive accuracy of both conventional models outstripped that of our AI-based methods suggests that future development of mixed-input CNN-LSTMs that analyze both raw ECG and demographic/co-morbidity data might prove useful.

Limitations: Our study has several additional limitations. First, the modest performance of our networks limits clinical applicability at the current stage. Although the accuracy we observed was not superior to the traditional model based on clinical data, we note that triage tools for potentially fatal diseases should prioritize high negative predictive value and our models did achieve that design goal. A deep learning model based on both clinical and ECG data might be able to achieve better accuracy; however, building such a network with heterogeneous inputs is complicated, and to the best of our knowledge, has not been attempted in this field. Notably, while we note that the size of the cohort considered in this study is small compared to many prior AI-based ECG studies,$^{39,40}$ our analysis (Supplemental Fig. 1) suggests this may not be the key factor constraining model performance. While this finding does not guarantee that increasing the cohort size would not increase accuracy of our model, it suggests a more likely explanation is that the deep learning approach is relatively ill-suited for the chosen problem.
Second, this study included only hospitalized patients with COVID-19. Whether this test can be used to prognosticate in settings where patients have few or no symptoms requires a separate study.

Third, certain agents used to treat COVID-19, which we did not aim to capture in our database, affect the ECG (e.g., hydroxychloroquine, azithromycin, and other critical care medications) and may have impacted network performance. Regarding the latter point, the geographic and management diversity of the included patients who were treated in different institutions may have enhanced the robustness and generalizability of our analysis.

Lastly, MACE definitions were based on clinicians’ determinations as adjudicated by chart review and notes. While we collected data on certain laboratory values, we opted to rely on clinician judgment for diagnosis of MACE elements such as acute MI. These definitions carry an important implication of subjectivity and inter-observer bias, but we determined that this would be the more ideal adjudication method compared to the alternative of having an arbitrary threshold lab value or PVC burden when these parameters were not being measured systematically in the pandemic setting.

Conclusions: Our analysis shows that AI-based algorithms could be helpful for rapid prognostication of adverse events in COVID-19 patients using a single ECG taken at the time of admission. While these models do not have perfect accuracy by any means, their performance is comparable to that of conventional statistical models relying on a multitude of clinical and demographic variables.
### Table 1: Demographics, comorbidities, and outcomes; overall and by enrolling hospital system.

| Demographics | Overall | UW | KI | UU | UC | P-value | Missing (%) |
|--------------|---------|----|----|----|----|---------|-------------|
| **N** | 1386 | 420 | 481 | 308 | 177 |
| **Age (years)** | | | | | | | |
| (mean (SD)) | 63.43 (16.90) | 63.26 (16.41) | 59.91 (17.04) | 65.88 (17.96) | 69.15 (13.24) | <0.001 | 0.0 |
| **Sex at birth = female (%)** | | | | | | | |
| | 547 (39.5) | 169 (40.2) | 161 (33.5) | 134 (43.5) | 83 (46.9) | 0.004 | 0.0 |
| **BMI (kg/m²)** | | | | | | | |
| (mean (SD)) | 28.38 (6.58) | 29.70 (7.81) | 28.63 (5.96) | 27.85 (6.90) | 27.06 (5.79) | 0.001 | 20.8 |
| **Ethnicity (%)** | | | | | | | |
| Hisp/Latinx | 85 (7.1) | 84 (20.2) | 0 (0.0) | 1 (0.3) | - | 13.1 |
| Non Hisp/Latinx | 609 (50.5) | 327 (78.6) | 13 (2.7) | 269 (78.3) | - | - |
| Unkn/Unavail | 511 (42.4) | 5 (1.2) | 468 (97.3) | 38 (12.3) | - | - |
| **Race (%)** | | | | | | | |
| FN/AK Native | 9 (0.6) | 9 (2.1) | 0 (0.0) | 0 (0.0) | - | <0.001 | 0.0 |
| Asian | 118 (8.5) | 71 (16.9) | 7 (1.5) | 40 (13.0) | - | <0.001 | 0.0 |
| Black/Afr Amer | 74 (5.3) | 57 (13.6) | 11 (2.3) | 6 (1.9) | - | <0.001 | 0.0 |
| HI FN/Pac Isl | 7 (0.5) | 7 (1.7) | 0 (0.0) | 0 (0.0) | - | 0.001 | 0.0 |
| White | 694 (50.1) | 265 (63.1) | 228 (47.4) | 201 (65.3) | - | <0.001 | 0.0 |
| Other | 57 (4.1) | 3 (0.7) | 50 (10.4) | 4 (1.3) | - | <0.001 | 0.0 |
| Unkn/Unavail | 252 (18.2) | 8 (1.9) | 187 (38.9) | 57 (18.5) | - | <0.001 | 0.0 |
| **Comorbidity** | | | | | | | |
| Hypertension (%) | 737 (54.0) | 214 (53.2) | 236 (49.3) | 184 (59.7) | 103 (58.2) | 0.021 | 1.4 |
| CAD (%) | 187 (13.8) | 59 (14.8) | 65 (13.7) | 48 (15.6) | 15 (8.5) | 0.144 | 2.1 |
| CIED (%) | - | - | - | - | - | 13.3 |
| Pacemaker | 22 (1.8) | 9 (2.2) | 8 (1.7) | 5 (1.6) | - | - |
| ICD | 5 (0.4) | 4 (1.0) | 1 (0.2) | 0 (0.0) | - | - |
| **Outcomes** | | | | | | | |
| Arrhythmic event (%) | 125 (9.0) | 28 (6.7) | 48 (10.0) | 36 (11.7) | 13 (7.3) | 0.084 | 0.0 |
| TE (%) | 132 (9.5) | 36 (8.6) | 52 (10.8) | 29 (9.4) | 15 (8.5) | 0.660 | 0.0 |
| HF (%) | 109 (7.9) | 19 (4.5) | 58 (12.1) | 23 (7.5) | 9 (5.1) | <0.001 | 0.0 |
| mortality (%) | 245 (17.7) | 88 (21.0) | 73 (15.2) | 69 (22.4) | 15 (8.5) | <0.001 | 0.0 |

P-values are for tests of differences between centers (continuous variables: ANOVA; categorical variables: chi-squared). UW = University of Washington; KI = Karolinska Institutet; UU = Uppsala University; UC = University of Copenhagen; BMI = body mass index; FN/AK native = First Nations or Alaskan native; HI FN = Hawaiian First Nations; CAD = coronary artery disease; CIED = cardiac implanted electronic device; ICD = implanted cardioverter defibrillator; TE = thromboembolic event; HF = heart failure.
Figure 1: Flowchart showing inclusion of patients into databases for both classification problems to be address via AI-based predictive modeling. The left branch of the tree concerns CNN-LSTM1, concerned with differentiating between ECGs from patients who survived vs. died. The right branch describes the database used for CNN-LSTM2, which independently predicts the likelihood that each ECG belongs to a patient from four groups (no event vs. MACE as shown in legend). ECG = electrocardiogram; AE = arrhythmic event; TE = thromboembolic event; HF = heart failure.
Figure 2: Schematics illustrating machine learning network architectures for CNN-LSTM1/2. Each network consists of three sections: (I) convolution layers, shown here as “feature maps” and “pooling” subsections; (II) recurrent neural network layers (labeled “Long-Short Term Memory”); and (III) fully connected layers that produce outputs (i.e., predicted probabilities for each class). ReLu = rectified linear unit; LSTM = long-short term memory; AE = arrhythmic event; TE = thromboembolic event; HF = heart failure.
Figure 3: Data summarizing predictive power of AI-based and conventional models. A: Receiver-operator characteristic (ROC) curves for CNN-LSTM1 and the corresponding conventional model, which attempted to differentiate between ECGs of patients who survived vs. died. B: False positive and false negative rates ($f_{pr}$ and $f_{nr}$, respectively) for CNN-LSTM1 as a function of model threshold; at a nominal threshold (dashed yellow line) associated with a 20% $f_{nr}$, the corresponding $f_{pr}$ (~70%) is shown by a dashed black arrow. C: ROC curves for CNN-LSTM2 (multi-label model independent prediction of different major adverse cardiovascular event types; single curve derived via macro-averaging) and the corresponding conventional model (binary prediction: any event vs. no event). See Fig. 4 and Supplemental Fig. 2 for additional plots. MACE = major adverse cardiovascular event; AUC = area under the curve.
Figure 4: Individual receiver operator characteristic (ROC) curves for the four independent classification tasks performed by CNN-LSTM2. A: Prediction of no event during hospitalization from intake ECG. B: Prediction of arrhythmic event. C: Prediction of thromboembolic event. D: Prediction of heart failure event. MACE = major adverse cardiovascular event; AUC = area under the curve.
Figure 5: Violin plots with box-and-whisker annotations showing raw network outputs. A: Results for classification task (survival vs. death) in CNN-LSTM1. B & C: Results for two classification tasks in CNN-LSTM2 (B: arrhythmic events; C: heart failure events). X-axis labels show ground truth labels; dashed lines in each panel show optimal classification thresholds, as explained in Methods. See Supplemental Fig. 3 for additional plots. AE = arrhythmic event; TE = thromboembolic event; HF = heart failure.
Figure 6: Summary data for holdout testing of both AI-based models trained with ECGs from before a cutoff date (June 9, 2020) and tested with ECGs from after that date. A: Receiver-operator characteristic (ROC) curve for CNN-LSTM1 trained and tested using the holdout protocol defined above. B: Same as A but for CNN-LSTM2. Both macro-averaged and individual event type ROCs are shown superimposed on this plot. MACE = major adverse cardiovascular event; AUC = area under the curve.
Supplemental Methods

We implemented our deep learning models using Python version 3.7.\textsuperscript{41} The models were developed using the Keras deep learning API\textsuperscript{42} with the Tensorflow backend\textsuperscript{43} and the Scikit-learn API.\textsuperscript{44,45} To enforce the assumption that all data points were independent and identically distributed random variables, data leakage was not allowed (i.e., for each iteration of network training, an individual patient’s ECG was only used in one set – training, validation, or testing). As part of the 10-fold cross-validation process, stratified data partitioning ensured that the training and testing sets always contained percentages of every target class consistent with proportions in the entire dataset. To avoid over- and under-fitting, an internal validation set was randomly split from the training set and used to evaluate the model over successive iterations.

ECG records obtained at 250 Hz were up-sampled to 500 Hz via interpolation. Standard 12-lead ECG records were recorded, but four leads (III, aVR, aVL, aVF) were omitted from analysis as these can be derived from linear combinations of other included leads (and as such represent redundant data). We used data preprocessing to help ensure useful features in raw signals were appropriately prioritized during network training. To facilitate comparison of ECGs in a standardized dataspace, we normalized all signals to the range [0,1]; normalization was carried out on a patient-by-patient basis, such that each individual lead was adjusted according to the global min/max of all eight leads.

The nature of the output layers for the two models (i.e., section III in Fig. 2) differed based on application. CNN-LSTM1 was a binary model (i.e., predicted likelihood of survival vs. death) where the last layer contained a standard “softmax” normalized exponential function; in contrast, CNN-LSTM2 was a multi-label model (independent likelihoods for all four event types: arrhythmic, heart failure, thromboembolic, or none) where the last layer was a sigmoid function. The sigmoid layer allowed for independent prediction of the likelihood of an ECG belonging to each possible class, which was necessary since some patients experienced multiple different MACE events (e.g., both arrhythmic and thromboembolic events).
During the learning process, we used Adaptive Moment Estimation (Adam) gradient descent,\textsuperscript{46} with categorical cross-entropy as a loss function for quantifying the error and iterative updating of weights and biases. Given the fact that our database was small by deep learning standards (n ≤ 1386 compared to, e.g., n ≈ 650k\textsuperscript{39}) and imbalanced (i.e., containing a much lower number of patients with compared to without events), we applied an early stopping callback to prevent overfitting. Effectively, this meant the training process was halted if the loss rate in the validation set had not improved for ten iterations.

Parameters for each network layer in CNN-LSTM1 and CNN-LSTM2 were set to optimize predictive power using hyper-parameter tuning. We ran the complete train/validate/test process for both networks using every possible permutation of a subset of seven parameters, identified the settings that optimized predictive power, then locked in that configuration. Parameters set via this process were: batch size; ECG normalization; ECG filtering to remove baseline wander; class-based weighting; and four aspects of model architecture (number of convolutional layers, number of LSTM layers, use of max-pooling layers).

\textbf{Supplemental Fig. 5} shows heatmaps for the entire hyper-parameter tuning process for CNN-LSTM1 (A) and CNN-LSTM2 (B); in each case, the optimal combination of seven parameters (indicated visually by asterisks) was selected by maximizing the mean AUROC across all 10 model iterations. Finalized network structure parameters for both cases are provided in \textbf{Supplemental Tables 4 and 5}.

To assess whether perturbation of CNN initial conditions might improve the performance of our networks, we implemented two types of transfer learning (TL) to test pre-training based on a different set of signals, drawn from the ECG PTB Diagnostic Database.\textsuperscript{47} TL was carried out as follows: 290 subjects were divided into training, validation and testing sets in a 8:1:1 ratio. All 549 records acquired from these subjects were then chopped into 10-second ECG strips. Each strip was labeled as either unhealthy (Myocardial infarction, Cardiomyopathy, Heart failure, Bundle branch block, Dysrhythmia, Myocardial hypertrophy, Valvular heart disease, Myocarditis, Miscellaneous) or healthy (none of the prior categories) based on meta-data for the corresponding subject. Using the best candidate model emerging from the hyper-parameter tuning process as our starting point, we trained a new instance of the CNN to differentiate between ECGs from healthy and unhealthy individuals using a set of 5724 unique ECG strips (training: 4462; validation: 517; testing: 745).
To test whether pretrained networks outperformed those with randomized initial conditions, we attempted two different TL approaches. In the “feature extraction” method (TL1), we froze the parameters of the pre-trained model then replaced the final layer with a fully connected network comprising layers with eight and two nodes. These new layers were then trained with our COVID-19 dataset as described elsewhere in Methods, while parameters elsewhere in the network were held constant at their pre-trained values. In the “fine-tuning” method (TL2), parameters in the pre-trained version were unfrozen and retrained with our COVID-19 data.
Supplemental Materials

**Supplemental Table 1:** Summary of atrial arrhythmias observed on 116 intake ECGs in the database.

Percentage values shown are for the fraction of the total cohort (1386 records).

| Arrhythmia Type                     | N (%)   |
|-------------------------------------|---------|
| Wandering atrial pacemaker         | 1 (0.07%)|
| Atrial fibrillation                 | 97 (7.0%)|
| Atrial flutter                      | 12 (0.87%)|
| Multifocal atrial tachycardia       | 5 (0.36%)|
| Other Supraventricular Tachycardia | 1 (0.07%)|
| All atrial arrhythmias              | 116 (8.4%)|
### Supplemental Table 2: Same as Table 1 but subdivided by survival vs. death.

|                         | death     | survival  | p-value | Missing (%) |
|-------------------------|-----------|-----------|---------|-------------|
| N                       | 245       | 1,141     |         |             |
| **Demographics**        |           |           |         |             |
| Age (years) (mean (SD)) | 74.46 (14.40) | 61.06 (16.46) | <0.001  | 0.0         |
| sex_at_birth = female (%) | 80 (32.7)     | 467 (40.9)     | 0.020   | 0.0         |
| BMI (kg/m$^2$) (mean (SD)) | 27.58 (6.91) | 28.54 (6.51) | 0.069   | 20.8        |
| Ethnicity (%)           |           |           | 0.014   | 13.1        |
| Hisp/Latinx             | 12 (5.2)  | 73 (7.5)  |         |             |
| Non Hisp/Latinx         | 136 (59.1) | 473 (48.5) |         |             |
| Unkn/Unavail            | 82 (35.7) | 429 (44.0) |         |             |
| **Race(%)**             |           |           |         |             |
| FN/AK Native            | 2 (0.8)   | 7 (0.6)   | 1.000   | 0.0         |
| Asian                   | 20 (8.2)  | 98 (8.6)  | 0.928   | 0.0         |
| Black/Afr Amer          | 13 (5.3)  | 61 (5.3)  | 1.000   | 0.0         |
| HI FN/Pac Isl           | 1 (0.4)   | 6 (0.5)   | 1.000   | 0.0         |
| White                   | 137 (55.9) | 557 (48.8) | 0.052   | 0.0         |
| Other                   | 11 (4.5)  | 46 (4.0)  | 0.880   | 0.0         |
| Unkn/Unavail            | 46 (18.8) | 206 (18.1) | 0.862   | 0.0         |
| **Comorbidity**         |           |           |         |             |
| Hypertension (%)        | 170 (70.2)| 567 (50.4) | <0.001  | 1.4         |
| Chronic Lung Disease (Chronic Obstructive Pulmonary Disease, Sarcoidosis, or other)(%) | 55 (22.9) | 162 (14.4) | 0.002 | 1.6 |
| Coronary artery disease (%) | 56 (23.4) | 131 (11.7) | <0.001 | 2.1 |
| Cardiomyopathy (%)      |           |           | <0.001  | 14.5        |
| Ischemic CM             | 29 (12.9) | 65 (6.8)  |         |             |
| Non-ischemic CM         | 12 (5.4)  | 28 (2.9)  |         |             |
| Mixed                   | 3 (1.3)   | 2 (0.2)   |         |             |
| No cardiomyopathy       | 180 (80.4)| 866 (90.1) |         |             |
| Pacer/ICD (%)           |           |           | 0.909   | 13.3        |
| Pacemaker               | 5 (2.2)   | 17 (1.7)  |         |             |
| ICD                     | 1 (0.4)   | 4 (0.4)   |         |             |
| Length of stay in ICU in days (mean (SD)) | 5.08 (11.09) | 3.14 (7.48) | 0.004 | 28.4 |

### Ventricular Arrhythmia
| Condition                                      | Control Group | Test Group | p-value | C-statistic |
|-----------------------------------------------|---------------|------------|---------|-------------|
| Premature Ventricular complexes (PVCs) (%)    | 13 (5.3)      | 18 (1.6)   | 0.001   | 0.0         |
| Non sustained ventricular tachycardia (NSVT) (%) | 2 (0.8)      | 6 (0.5)    | 0.936   | 0.0         |
| Sustained Ventricular Tachycardia (VT) (%)    | 4 (1.6)       | 1 (0.1)    | 0.002   | 0.0         |
| Ventricular fibrillation (VF) (%)             | 3 (1.2)       | 1 (0.1)    | 0.019   | 0.0         |
| NONE (%)                                      | 212 (86.5)    | 953 (83.5) | 0.284   | 0.0         |
| Atrial Fibrillation (%)                       | <0.001        | 0.0        |         |             |
| No AF                                         | 169 (69.0)    | 1005 (88.1)|         |             |
| Recurrent Episode of Known Paroxysmal AF      | 8 (3.3)       | 26 (2.3)   |         |             |
| Persistent/Permanent AF                       | 32 (13.1)     | 58 (5.1)   |         |             |
| New Onset AF                                  | 36 (14.7)     | 52 (4.6)   |         |             |
| **Major Adverse Cardiovascular Event**        |               |            |         |             |
| Cardiogenic Shock (%)                         | 10 (4.1)      | 3 (0.3)    | <0.001  | 0.0         |
| Heart Failure (%)                             | 49 (20.0)     | 54 (4.7)   | <0.001  | 0.0         |
| Sustained Ventricular Arrhythmia (%)          | 3 (1.2)       | 1 (0.1)    | 0.019   | 0.0         |
| Tachyarrhythmia Cardiac Arrest (%)            | 4 (1.6)       | 5 (0.4)    | 0.094   | 0.0         |
| Bradyarrhythmia Cardiac Arrest (%)            | 8 (3.3)       | 2 (0.2)    | <0.001  | 0.0         |
| Acute Coronary Syndrome (%)                   | 8 (3.3)       | 10 (0.9)   | 0.007   | 0.0         |
| Pericardial Tamponade (%)                     | 0 (0.0)       | 1 (0.1)    | 1.000   | 0.0         |
| Thromboembolic Complication (%)               | 36 (14.7)     | 96 (8.4)   | 0.004   | 0.0         |
| NONE (%)                                      | 141 (57.6)    | 847 (74.2) | <0.001  | 0.0         |
| **Overall Outcomes**                          |               |            |         |             |
| Died (%)                                      | 245 (100.0)   | 0 (0.0)    | <0.001  | 0.0         |
| Discharged to Comfort care OR Hospice (%)     | 7 (2.9)       | 27 (2.4)   | 0.824   | 0.0         |
| Discharged home (%)                           | 0 (0.0)       | 893 (78.3) | <0.001  | 0.0         |
| Discharged to rehabilitation unit OR different speciality OR lower care unit after COVID stabilization (%) | 1 (0.4) | 224 (19.6) | <0.001 | 0.0         |

### Automatically Extracted ECG Metrics (mean ± standard deviation)

| Metric                          | Control Group | Test Group | p-value | C-statistic |
|---------------------------------|---------------|------------|---------|-------------|
| Ventricular Rate (beats per min.) | 95.1 ± 22.1  | 90.1 ± 19.1| 0.0017  | 1.0         |
| PR interval (ms)                | 159.1 ± 34.1  | 155.1 ± 28.1| 0.1347  | 9.7         |
| QRS duration (ms)               | 98.1 ± 20.1   | 93.1 ± 18.1| 0.0004  | 1.0         |
| Corrected QT interval (ms)      | 454.1 ± 36.1  | 444.1 ± 31.1| <0.0001 | 1.0         |
| P-wave duration (ms)            | 51.1 ± 9.1    | 53.1 ± 7.1 | 0.1209  | 35.2        |
Supplemental Table 3: Same as Table 1 but subdivided by cardiovascular event type.

|                | None     | Arrhythmic | TE       | HF       | P-value | Missing (%) |
|----------------|----------|------------|----------|----------|---------|-------------|
| N              | 1,090    | 119        | 130      | 107      |         |             |

**Demographics**

|                                | None     | Arrhythmic | TE       | HF       | P-value | Missing (%) |
|--------------------------------|----------|------------|----------|----------|---------|-------------|
| Age (years) (mean (SD))        | 62.17 (17.09) | 69.13 (15.57) | 62.74 (14.11) | 72.20 (14.96) | <0.001 | 0.0          |
| sex_at_birth = female (%)      | 452 (41.5) | 38 (31.9)  | 39 (30.0) | 32 (29.9) | 0.004  | 0.0          |
| BMI (kg/m²) (mean (SD))        | 28.24 (6.44) | 28.37 (6.65) | 28.09 (6.26) | 29.41 (7.55) | 0.406  | 20.5         |

**Ethnicity (%)**

|                | None     | Arrhythmic | TE       | HF       | P-value | Missing (%) |
|----------------|----------|------------|----------|----------|---------|-------------|
| Hisp/Latinx     | 74 (7.9) | 3 (2.8)    | 9 (7.8)  | 1 (1.0)  |         |             |
| Non Hisp/Latinx | 492 (52.3) | 55 (51.9)  | 50 (43.5) | 33 (33.7) |         |             |
| Unkn/Unavail    | 375 (39.9) | 48 (45.3)  | 56 (48.7) | 64 (65.3) |         |             |

**Race (%)**

|                | None     | Arrhythmic | TE       | HF       | P-value | Missing (%) |
|----------------|----------|------------|----------|----------|---------|-------------|
| FN/AK Native   | 6 (0.6)  | 1 (0.8)    | 2 (1.5)  | 1 (0.9)  | 0.614  | 0.0          |
| Asian          | 99 (9.1) | 5 (4.2)    | 8 (6.2)  | 6 (5.6)  | 0.148  | 0.0          |
| Black/Afr Amer | 63 (5.8) | 5 (4.2)    | 7 (5.4)  | 2 (1.9)  | 0.350  | 0.0          |
| HI FN/Pac Isl  | 7 (0.6)  | 0 (0.0)    | 0 (0.0)  | 0 (0.0)  | 0.513  | 0.0          |
| White          | 531 (48.7) | 67 (56.3)  | 73 (56.2) | 58 (54.2) | 0.155  | 0.0          |
| Other          | 38 (3.5) | 9 (7.6)    | 6 (4.6)  | 6 (5.6)  | 0.145  | 0.0          |
| Unkn/Unavail   | 202 (18.5) | 20 (16.8)  | 19 (14.6) | 25 (23.4) | 0.360  | 0.0          |

**Comorbidity**

|                                | None     | Arrhythmic | TE       | HF       | P-value | Missing (%) |
|--------------------------------|----------|------------|----------|----------|---------|-------------|
| Hypertension (%)               | 544 (50.7) | 80 (68.4)  | 60 (47.2) | 88 (83.0) | <0.001 | 1.5          |
| Chronic Lung Disease (Chronic Obstructive Pulmonary Disease, Sarcoidosis, or other)(%) | 159 (14.8) | 25 (21.4)  | 20 (15.7) | 27 (25.7) | 0.013  | 1.7          |
| Coronary artery disease (%)    | 126 (11.8) | 26 (22.4)  | 13 (10.2) | 35 (33.0) | <0.001 | 2.1          |
| Cardiomyopathy (%)             |         |            |          |          | <0.001 | 14.4         |
| Ischemic CM                    | 58 (6.3) | 16 (15.4)  | 7 (6.2)  | 22 (22.7) |         |             |
| Non-ischemic CM                | 23 (2.5) | 4 (3.8)    | 4 (3.6)  | 12 (12.4) |         |             |
| Mixed                          | 4 (0.4) | 0 (0.0)    | 0 (0.0)  | 1 (1.0)  |         |             |
| No cardiomyopathy              | 840 (90.8) | 84 (80.8)  | 101 (90.2) | 62 (63.9) |         |             |
| Pacer/ICD (%)                  |         |            |          |          | 0.110  | 13.1         |
| Pacemaker                      | 15 (1.6) | 2 (1.9)    | 0 (0.0)  | 5 (5.2)  |         |             |
| ICD                                      | 5 ( 0.5) | 0 ( 0.0) | 0 ( 0.0) | 0 ( 0.0) |
|------------------------------------------|----------|----------|----------|----------|
| **Length of stay in ICU in days (mean (SD))** | 2.26 (5.93) | 10.32 (15.54) | 10.33 (13.92) | 6.71 (11.79) | <0.001 | 28.2 |
| **Ventricular Arrhythmia**                |          |          |          |          |
| Premature Ventricular complexes (PVCs) (%) | 0 ( 0.0) | 31 (26.1) | 10 ( 7.7) | 9 ( 8.4) | <0.001 | 0.0 |
| Non sustained ventricular tachycardia (NSVT) (%) | 0 ( 0.0) | 4 ( 3.4) | 1 ( 0.8) | 4 ( 3.7) | <0.001 | 0.0 |
| Sustained Ventricular Tachycardia (VT) (%) | 0 ( 0.0) | 5 ( 4.2) | 1 ( 0.8) | 3 ( 2.8) | <0.001 | 0.0 |
| Ventricular fibrillation (VF) (%)         | 0 ( 0.0) | 4 ( 3.4) | 1 ( 0.8) | 1 ( 0.9) | <0.001 | 0.0 |
| NONE (%)                                 | 945 (86.7) | 66 (55.5) | 102 (78.5) | 85 (79.4) | <0.001 | 0.0 |
| **Atrial Fibrillation (%)**               |          |          |          |          | <0.001 | 0.0 |
| No AF                                    | 1002 (91.9) | 33 (27.7) | 110 (84.6) | 56 (52.3) |
| Recurrent Episode of Known Paroxysmal AF  | 25 (2.3) | 0 ( 0.0) | 2 ( 1.5) | 8 ( 7.5) |
| Persistent/Permanent AF                  | 63 (5.8) | 4 ( 3.4) | 4 ( 3.1) | 22 (20.6) |
| New Onset AF                             | 0 ( 0.0) | 82 (68.9) | 14 (10.8) | 21 (19.6) |
| **Major Adverse Cardiovascular Event**    |          |          |          |          |          |          |
| Cardiogenic Shock (%)                    | 0 ( 0.0) | 5 ( 4.2) | 3 ( 2.3) | 13 (12.1) | <0.001 | 0.0 |
| Heart Failure (%)                        | 0 ( 0.0) | 30 (25.2) | 12 ( 9.2) | 101 (94.4) | <0.001 | 0.0 |
| Sustained Ventricular Arrhythmia (%)     | 0 ( 0.0) | 4 ( 3.4) | 1 ( 0.8) | 1 ( 0.9) | <0.001 | 0.0 |
| Tachyarrhythmia Cardiac Arrest (%)       | 0 ( 0.0) | 9 ( 7.6) | 3 ( 2.3) | 2 ( 1.9) | <0.001 | 0.0 |
| Bradyarrhythmia Cardiac Arrest (%)       | 0 ( 0.0) | 10 ( 8.4) | 3 ( 2.3) | 3 ( 2.8) | <0.001 | 0.0 |
| Acute Coronary Syndrome (%)              | 7 ( 0.6) | 3 ( 2.5) | 3 ( 2.3) | 8 ( 7.5) | <0.001 | 0.0 |
| Pericardial Tamponade (%)                | 0 ( 0.0) | 1 ( 0.8) | 1 ( 0.8) | 0 ( 0.0) | 0.022 | 0.0 |
| Thromboembolic Complication (%)          | 0 ( 0.0) | 27 (22.7) | 130 (100.0) | 13 (12.1) | <0.001 | 0.0 |
| NONE (%)                                 | 937 (86.0) | 48 (40.3) | 0 ( 0.0) | 0 ( 0.0) | <0.001 | 0.0 |
| **Overall Outcomes**                     |          |          |          |          |          |          |
| Died (%)                                 | 134 (12.3) | 54 (45.4) | 35 (26.9) | 53 (49.5) | <0.001 | 0.0 |
| Discharged to Comfort care OR Hospice (%) | 26 (2.4) | 3 ( 2.5) | 1 ( 0.8) | 4 ( 3.7) | 0.506 | 0.0 |
| Discharged home (%) | 792 (72.7) | 35 (29.4) | 55 (42.3) | 22 (20.6) | <0.001 | 0.0 |
|---------------------|------------|----------|----------|----------|--------|-----|
| Discharged to rehabilitation unit OR different speciality OR lower care unit after COVID stabilization (%) | 146 (13.4) | 29 (24.4) | 40 (30.8) | 28 (26.2) | <0.001 | 0.0 |

Automatically Extracted ECG Metrics (mean ± standard deviation)

| Metric                          | Mean ± Standard Deviation | Mean ± Standard Deviation | Mean ± Standard Deviation | Mean ± Standard Deviation | p value | Effect Size |
|---------------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------|-------------|
| Ventricular Rate (bpm)          | 89.1±18.1                 | 90.1±19.1                 | 98.1±22.1                 | 95.1±23.1                 | <0.0001 | 1.1         |
| PR interval (ms)                | 155.1±27.1                | 157.1±30.1                | 156.1±38.1                | 158.1±35.1                | 0.7585  | 11.4        |
| QRS duration (ms)               | 93.1±18.1                 | 93.1±16.1                 | 104.1±22.1                | 97.1±19.1                 | <0.0001 | 1.1         |
| Corr. QT interval (ms)          | 443.1±31.1                | 451.1±29.1                | 457.1±38.1                | 454.1±33.1                | <0.0001 | 1.1         |
| P-wave duration (ms)            | 53.1±7.1                  | 54.1±7.1                  | 51.1±13.1                 | 52.1±9.1                  | 0.4717  | 37.6        |
Supplemental Table 3a: Average covariates from the 10 conventional logistic regression models with the outcomes of Mortality and Combined Cardiac Event. Covariates are based on the train datasets.

| Covariate                          | Mortality vs. Survival | Combined Cardiac Event |
|-----------------------------------|------------------------|------------------------|
| Covariate                         | Odds Ratio  | 95% CI   | P-value | Odds Ratio  | 95% CI   | P-value |
| Age                               | 1.07       | 1.05, 1.08 | <0.001  | 1.02       | 1.01, 1.03 | <0.01   |
| Race (ref = White, Non-Hispanic)  |            |          |         |            |          |         |
| Asian                             | 1.03       | 0.54, 1.97 | 0.78    | 0.76       | 0.38, 1.53 | 0.49    |
| Black                             | 0.98       | 0.45, 2.15 | 0.85    | 0.73       | 0.32, 1.66 | 0.48    |
| Other                             | 0.73       | 0.50, 0.38 | 0.11    | 1.42       | 1.00, 2.03 | 0.064   |
| White, Hispanic                   | 0.92       | 0.38, 2.22 | 0.77    | 0.83       | 0.37, 1.87 | 0.69    |
| BMI                               | 1.02       | 0.99, 1.05 | 0.20    | 1.03       | 1.00, 1.06 | 0.030   |
| Sex                               | 0.56       | 0.38, 0.81 | <0.01   | 0.65       | 0.47, 0.91 | 0.015   |
| Hypertension                      | 1.05       | 0.71, 1.54 | 0.68    | 1.33       | 0.94, 1.89 | 0.13    |
| Coronary Artery Disease           | 1.31       | 0.84, 2.04 | 0.28    | 1.42       | 0.92, 2.17 | 0.17    |
| Pacer                             | 0.63       | 0.20, 2.02 | 0.49    | 0.87       | 0.30, 2.57 | 0.73    |
| ECG features                      |            |          |         |            |          |         |
| Ventricular Rate                  | 1.02       | 1.01, 1.02 | <0.01   | 1.01       | 1.01, 1.02 | <0.01   |
| PR interval                       | 1.00       | 1.00, 1.00 | 0.56    | 1.00       | 0.99, 1.00 | 0.034   |
| QRS duration                      | 1.00       | 0.99, 0.99 | 0.46    | 1.00       | 0.99, 1.01 | 0.53    |
| Corr. QT interval                 | 1.00       | 0.99, 1.00 | 0.31    | 1.00       | 1.00, 1.00 | 0.61    |
### Supplemental Table 4: Empirically optimized parameters for convolution and LSTM layers in CNN-LSTM1.

| Layer     | Output Shape | Pool size | Kernel size | Number of filters | strides | Number of LSTM units |
|-----------|--------------|-----------|-------------|-------------------|---------|----------------------|
| Input     | 5000x8       | -         | -           | -                 | -       | -                    |
| Conv1d_1  | 2497x16      | -         | 7           | 16                | 2       | -                    |
| Conv1d_2  | 1246x16      | -         | 7           | 16                | 2       | -                    |
| MaxPooling| 623x16       | 2         | -           | -                 | 2       | -                    |
| Conv1d_3  | 310x32       | -         | 5           | 32                | 2       | -                    |
| Conv1d_4  | 153x32       | -         | 5           | 32                | 2       | -                    |
| MaxPooling| 76x32        | 2         | -           | -                 | -       | -                    |
| LSTM_1    | 76x4         | -         | -           | -                 | -       | 4                    |
| LSTM_2    | 76x4         | -         | -           | -                 | -       | 4                    |
| Flatten   | 304          | -         | -           | -                 | -       | -                    |
| Dense (Softmax) | 2 | - | - | - | - | - |

### Supplemental Table 5: Empirically optimized parameters for convolution and LSTM layers in CNN-LSTM2.

| Layer     | Output Shape | Kernel size | Number of filters | strides | Number of LSTM units |
|-----------|--------------|-------------|-------------------|---------|----------------------|
| Input     | 5000x8       | -           | -                 | -       | -                    |
| Conv1d_1  | 2497x16      | 7           | 16                | 2       | -                    |
| Conv1d_2  | 1246x16      | 7           | 16                | 2       | -                    |
| Conv1d_3  | 621x32       | 5           | 32                | 2       | -                    |
| Conv1d_4  | 309x32       | 5           | 32                | 2       | -                    |
| LSTM_1    | 309x4        | -           | -                 | -       | 4                    |
| LSTM_2    | 309x4        | -           | -                 | -       | 4                    |
| Flatten   | 1236         | -           | -                 | -       | -                    |
| Dense (Sigmoid) | 4 | - | - | - | - | - |
Supplemental Figures

Supplemental Figure 1: Learning curve showing the relationship between model accuracy \((\frac{tp + tn}{N})\) and cohort size for CNN-LSTM1, as described in Methods. Cohort sizes used in this analysis had 20%, 40%, 60%, or 80% of the total population (i.e., n = 277, 556, 835, of 1114 of 1386 records); we ensured that the ratio of records for patients who survived vs. those who died remained static in all four sub-sets.
Supplemental Figure 2: Plots of false positive and false negative rates ($fpr$ and $fnr$, respectively) for different label types in CNN-LSTM2 as a function of model threshold. Same panel ordering as in Fig. 4. See Fig. 3B legend for complete information about interpreting these plots and annotations.
Supplemental Figure 3: Additional violin plots with box-and-whisker annotations showing raw network outputs. See legend for Fig. 5 for further information about graph labels, etc. A: Prediction of no event label. B: Prediction of thromboembolic events. AE = arrhythmic event; TE = thromboembolic event; HF = heart failure.

Supplemental Figure 4: Temporal distribution of patient admissions, mortality rate, and MACE incidence. A: Overall recruitments over the 12 months of the study. B: Proportion of patients recruited by each center over time. C&D: Proportion of patients who died or experienced MACE, organized by month of admission. UW = University of Washington; KI = Karolinska Institutet; UU = Uppsala University; UC = University of Copenhagen.
Supplemental Figure 5: Heatmaps of mean AUROC values for all 128 combinations of hyper-parameters used in the tuning process for CNN-LSTM1 (A) and CNN-LSTM2 (B). Asterisks indicate the optimal parameters combinations for each network (as outlined in Supplemental Methods), which were selected in each case by maximizing mean AUROC across all 10 model cross-validation iterations.

References

1. Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J. & Prescott, H. C. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. JAMA 324, 782-793, doi:10.1001/jama.2020.12839 (2020).

2. World Health Organization. WHO Coronavirus (COVID-19) Dashboard, <https://covid19.who.int> (2021).

3. Garcia, S. et al. Reduction in ST-Segment Elevation Cardiac Catheterization Laboratory Activations in the United States During COVID-19 Pandemic. J Am Coll Cardiol 75, 2871-2872, doi:10.1016/j.jacc.2020.04.011 (2020).

4. Sharma, R. et al. Excess Cerebrovascular Mortality in the United States During the COVID-19 Pandemic. Stroke 52, 563-572, doi:10.1161/STROKEAHA.120.031975 (2021).

5. Kaufman, H. W., Chen, Z., Niles, J. & Fesko, Y. Changes in the Number of US Patients With Newly Identified Cancer Before and During the Coronavirus Disease 2019 (COVID-19) Pandemic. JAMA Netw Open 3, e2017267, doi:10.1001/jamanetworkopen.2020.17267 (2020).

6. Blumenthal, D., Fowler, E. J., Abrams, M. & Collins, S. R. Covid-19 - Implications for the Health Care System. N Engl J Med 383, 1483-1488, doi:10.1056/NEJMc2021088 (2020).

7. Nishiga, M., Wang, D. W., Han, Y., Lewis, D. B. & Wu, J. C. COVID-19 and cardiovascular disease: from basic mechanisms to clinical perspectives. Nat Rev Cardiol 17, 543-558, doi:10.1038/s41569-020-0413-9 (2020).

8. Marchiano, S. et al. SARS-CoV-2 Infects Human Pluripotent Stem Cell-Derived Cardiomyocytes, Impairing Electrical and Mechanical Function. Stem Cell Reports 16, 478-492, doi:10.1016/j.stemcr.2021.02.008 (2021).
Michaud, G. F. & Stevenson, W. G. Atrial Fibrillation. *N Engl J Med* **384**, 353-361, doi:10.1056/NEJMcp2023658 (2021).

Shi, S. *et al.* Association of Cardiac Injury With Mortality in Hospitalized Patients With COVID-19 in Wuhan, China. *JAMA Cardiol* **5**, 802-810, doi:10.1001/jamacardio.2020.0950 (2020).

Castro, R. A. & Frishman, W. H. Thrombotic Complications of COVID-19 Infection: A Review. *Cardiol Rev* **29**, 43-47, doi:10.1097/CRD.0000000000000347 (2021).

Modin, D. *et al.* Acute COVID-19 and the Incidence of Ischemic Stroke and Acute Myocardial Infarction. *Circulation* **142**, 2080-2082, doi:10.1161/CIRCULATIONAHA.120.050809 (2020).

Katsoularis, I., Fonseca-Rodriguez, O., Farrington, P., Lindmark, K. & Fors Connolly, A. M. Risk of acute myocardial infarction and ischaemic stroke following COVID-19 in Sweden: a self-controlled case series and matched cohort study. *Lancet* **398**, 599-607, doi:10.1016/S0140-6736(21)00896-5 (2021).

Fye, W. B. A history of the origin, evolution, and impact of electrocardiography. *Am J Cardiol* **73**, 937-949, doi:10.1016/0002-9149(94)90135-x (1994).

Lanza, G. A. *et al.* Electrocardiographic findings at presentation and clinical outcome in patients with SARS-CoV-2 infection. *EP Europace* **23**, 123-129, doi:10.1093/eurpace/eeu245 (2021).

Christopoulos, G. *et al.* Artificial Intelligence-Electrocardiography to Predict Incident Atrial Fibrillation: A Population-Based Study. *Circ Arrhythm Electrophysiol* **13**, e009355, doi:10.1161/CIRCEP.120.009355 (2020).

Ko, W. Y. *et al.* Detection of Hypertrophic Cardiomyopathy Using a Convolutional Neural Network-Enabled Electrocardiogram. *J Am Coll Cardiol* **75**, 722-733, doi:10.1016/j.jacc.2019.12.030 (2020).

Adedinsewo, D. *et al.* Artificial Intelligence-Enabled ECG Algorithm to Identify Patients With Left Ventricular Systolic Dysfunction Presenting to the Emergency Department With Dyspnea. *Circ Arrhythm Electrophysiol* **13**, e008437, doi:10.1161/CIRCEP.120.008437 (2020).

Kwon, J. M. *et al.* Deep Learning-Based Algorithm for Detecting Aortic Stenosis Using Electrocardiography. *J Am Heart Assoc* **9**, e014717, doi:10.1161/JAHA.119.014717 (2020).

Raghunath, S. *et al.* Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med* **26**, 886-891, doi:10.1038/s41591-020-0870-z (2020).

Kwon, J. M. *et al.* Artificial intelligence algorithm for predicting cardiac arrest using electrocardiography. *Scand J Trauma Resusc Emerg Med* **28**, 98, doi:10.1186/s13049-020-00791-0 (2020).

LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).

Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput* **9**, 1735-1780, doi:10.1162/neco.1997.9.8.1735 (1997).

Harris, P. A. *et al.* Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* **42**, 377-381, doi:10.1016/j.jbi.2008.08.010 (2009).
Harris, P. A. et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform* **95**, 103208, doi:10.1016/j.jbi.2019.103208 (2019).

Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry* **77**, 534-540, doi:10.1001/jamapsychiatry.2019.3671 (2020).

Nair, V. & Hinton, G. E. in *Proceedings of the 27th International Conference on International Conference on Machine Learning* 807–814 (Omnipress, Haifa, Israel, 2010).

Perlich, C., Provost, F. & Simonoff, J. Tree induction vs. logistic regression: A learning-curve analysis. *J Mach Learn Res* **4**, 211--255 (2003).

CRAN - Package readxl v. 1.3.1 (Comprehensive R Archive Network (CRAN), 2021).

CRAN - Package tidyr v. 1.1.3 (Comprehensive R Archive Network (CRAN), 2021).

CRAN - Package dplyr v. 1.0.7 (Comprehensive R Archive Network (CRAN), 2021).

CRAN - Package caret v. 6.0-86 (Comprehensive R Archive Network (CRAN), 2021).

Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77, doi:10.1186/1471-2105-12-77 (2011).

De Vita, A. et al. Electrocardiographic Findings and Clinical Outcome in Patients with COVID-19 or Other Acute Infectious Respiratory Diseases. *Journal of Clinical Medicine* **9**, 3647, doi:10.3390/jcm9113647 (2020).

Sanders, G. D. et al. *Comparative Effectiveness Review* (ed U.S. Department of Health and Human Services) (Agency for Healthcare Research and Quality, Rockville, MD, 2018).

Wiley, Z. et al. Racial and Ethnic Differences and Clinical Outcomes of COVID-19 Patients Presenting to the Emergency Department. *Clin Infect Dis*, doi:10.1093/cid/ciab290 (2021).

Jambukia, S. H., Dabhi, V. K. & Prajapati, H. B. in *2015 International Conference on Advances in Computer Engineering and Applications*. 714-721.

Hicks, S. A. et al. Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. *Sci Rep* **11**, 10949, doi:10.1038/s41598-021-90285-5 (2021).

Attia, Z. I. et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* **394**, 861-867, doi:10.1016/S0140-6736(19)31721-0 (2019).

Attia, Z. I. et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* **25**, 70-74, doi:10.1038/s41591-018-0240-2 (2019).

Rossum, G. v. & Boer, J. d. Interactively Testing Remote Servers Using the Python Programming Language. *CWI Quarterly* **4** (1991).

Keras (GitHub, GitHub repository, 2015).

TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (2015).

CVDHJ-D-21-00058 • Revised Manuscript • Sridhar et al.
Al-Based Prediction of Adverse COVID-19 Outcomes from Intake ECGs
Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825--2830 (2011).

Buitinck, L. et al. API design for machine learning software: experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108--122 (2013).

Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Bousseljot, R., Kreiseler, D. & Schnabel, A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. (1995).
Key Findings

- We demonstrate the feasibility and examine the potential effectiveness of a rapid, inexpensive triage tool for COVID-19 patients using a 12-lead ECG at admission augmented by machine learning.
- When compared with the traditional statistical model incorporating extensive clinical data, our results using only the DNN-augmented ECG did not yield improved discrimination.
- Our models' low sensitivity and positive predictive value would be a barrier to successful deployment in clinical settings in its current form; our findings suggest that mixed-input models using both ECG data and clinical variables might be a viable path towards better predictive capabilities.