JACC: ADVANCES VOL. 1, NO. 1, 2022

ORIGINAL RESEARCH

EMERGING TECHNOLOGIES AND INNOVATIONS

A Deep Learning Model for Inferring Elevated Pulmonary Capillary Wedge Pressures From the 12-Lead Electrocardiogram

Daphne E. Schlesinger, BS,[a,b,c] Nathaniel Diamant, BS,[d] Aniruddh Raghu, MEng,[c,e] Erik Reinertsen, MD, PhD,[c,f] Katherine Young, MEng,[a] Puneet Batra, PhD,[d] Eugene Pomerantsev, MD, PhD,[f] Collin M. Stultz, MD, PhD[a,b,c,e,f]

ABSTRACT

BACKGROUND Central hemodynamic parameters are typically measured via pulmonary artery catherization—an invasive procedure that involves some risk to the patient and is not routinely available in all settings.

OBJECTIVES This study sought to develop a noninvasive method to identify elevated mean pulmonary capillary wedge pressure (mPCWP).

METHODS We leveraged data from 248,955 clinical records at the Massachusetts General Hospital to develop a deep learning model that can infer when the mPCWP >15 mmHg using the 12-lead electrocardiogram (ECG). Of these data, 242,216 records were used to pre-train a model that generates useful ECG representations. The remaining 6,739 records contain encounters with direct measurements of the mPCWP. Eighty percent of these data were used for model development and testing (5,390), and the remaining records comprise a holdout set (1,349) that was used to evaluate the model. We developed an associated unreliability score that identifies when model predictions are likely to be untrustworthy.

RESULTS The model achieves an area under the receiver operating characteristic curve (AUC) of 0.80 ± 0.02 (test set) and 0.79 ± 0.01 (holdout set). Model performance varies as a function of the unreliability, where patients with high unreliability scores correspond to a subgroup where model performance is poor: for example, patients in the holdout set with unreliability scores in the highest decile have a reduced AUC of 0.70 ± 0.06.

CONCLUSIONS The mPCWP can be inferred from the ECG, and the reliability of this inference can be measured. When invasive monitoring cannot be expeditiously performed, deep learning models may provide information that can inform clinical care. (JACC Adv 2022;1:100003) © 2022 The Authors. Published by Elsevier on behalf of the American College of Cardiology Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

From the [a]Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts, USA; [b]Institute for Medical Engineering and Science, MIT, Cambridge, Massachusetts, USA; [c]Research Laboratory of Electronics, Computer Science & Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts, USA; [d]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; [e]Department of Electrical Engineering and Computer Science, MIT, Cambridge, Massachusetts, USA; and the [f]Division of Cardiology, Massachusetts General Hospital, Boston, Massachusetts, USA.

The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

Manuscript received November 16, 2021; revised manuscript received December 28, 2021, accepted January 19, 2022.

ISSN 2772-963X https://doi.org/10.1016/j.jacadv.2022.100003

ABBREVIATIONS AND ACRONYMS

ECG = electrocardiogram

HF = heart failure

mPCWP = mean pulmonary capillary wedge pressure

RHC = right heart catheterization

Although standard hemodynamic parameters such as blood pressure and heart rate are readily obtained by the bedside, central hemodynamic parameters are challenging to infer from the physical examination alone and are only reliably measured via the insertion of a pulmonary artery catheter (PAC).[1,2] While PAC-guided care has not been shown to reduce mortality or length of hospital stay in critically ill patients, measurements of advanced hemodynamic parameters have important diagnostic and therapeutic implications in a variety of patient cohorts.[3-6] The mean pulmonary capillary wedge pressure (mPCWP), for example, is a strong predictor of post-discharge outcomes in patients admitted with severe symptomatic heart failure (HF), and hemodynamic congestion with concomitant elevations in left-sided pressures often precedes the onset of symptoms.[7-10]

The gold standard procedure for measuring advanced hemodynamic parameters, PAC insertion, is associated with complications, ranging from benign self-limited arrhythmias to rare, but often fatal, pulmonary artery perforations.[11,12] The procedure cannot always be scheduled expeditiously, especially in the setting of a global pandemic. Noninvasive methods that infer when important hemodynamic parameters are abnormal could therefore guide clinical decisions when an invasive procedure cannot be performed, or in the context of telemedicine. To these ends, we postulated that deep learning could be leveraged to estimate when the mPCWP is elevated using a readily available and routinely acquired signal, the electrocardiogram (ECG).

Deep learning is a subfield of machine learning where complex models are used to learn from data. Traditional models that have been used for clinical inference (eg, logistic regression and Cox proportional hazards survival models) typically require one to hand-pick features that are related to the task of interest. For example, the Pooled Cohort Equations utilize a predefined set of prognostic features that are related to major adverse cardiovascular events.[13] By contrast, deep learning typically takes raw data as input and therefore does not require the user to make decisions with respect to the features that are most informative. Such methods can therefore be useful when it is not initially clear what features from the data will be most dispositive. Classic examples include deep learning applied to medical images, where models are constructed that take all of the pixels within an image as input.[14] In the present study, we use a type of deep neural network, a convolutional neural network, which takes the entire 12-lead ECG as input and infers the mPCWP.

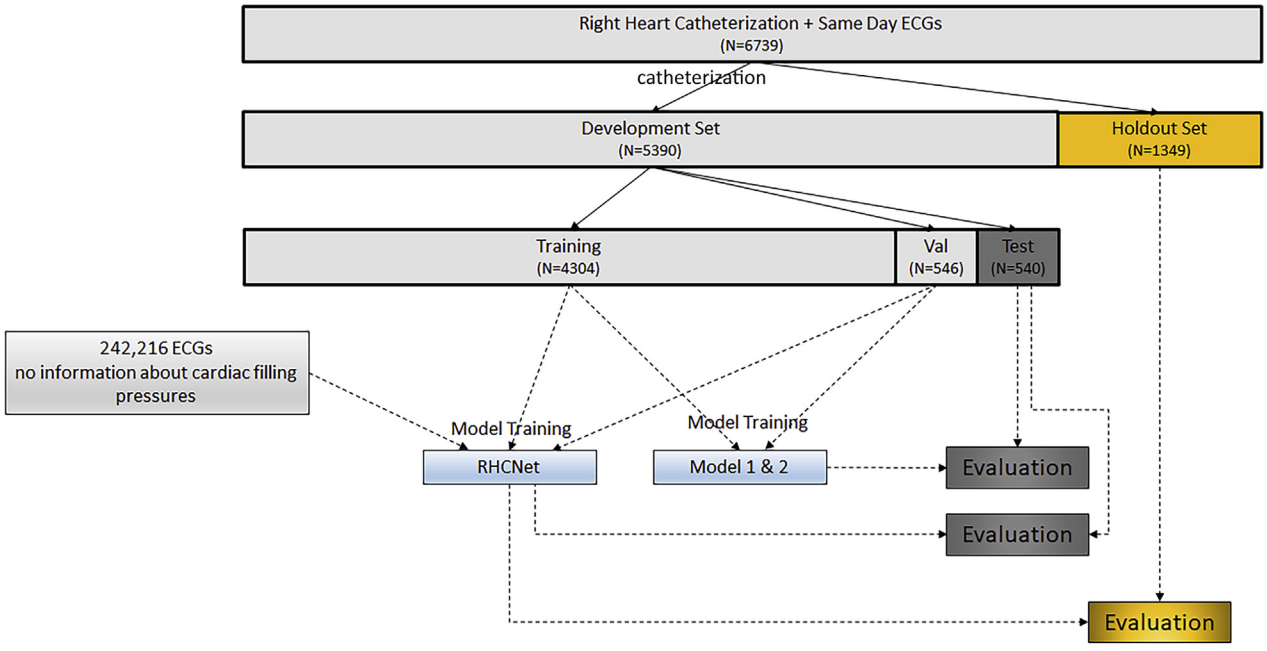
Deep learning algorithms have been applied to the 12-lead ECG for a variety of predictive tasks. Although many of these approaches have demonstrated remarkable discriminatory ability for problems that seem unrelated to the electrical activity of the heart,[15-17] studies that use the ECG to identify abnormal hemodynamic values are sparse. Three recent studies developed deep neural networks to identify patients with pulmonary hypertension (PH), yielding models with impressive discriminatory ability.[15,18,19] In these studies, the diagnosis of PH was made using echocardiography—that is, the mean velocity of a regurgitant jet across the tricuspid valve is directly related to the mean pulmonary arterial pressure.[20] While one can obtain information about the mPCWP using cardiac ultrasound, such insights require precise spectral Doppler measurements of mitral inflow velocities[21]—a task that is not within the average health care provider's skillset. These approaches can also be unreliable in patients who do not have echocardiographic evidence of diastolic dysfunction.[21] The reliable identification of elevated mPCWP, using readily available clinical data, remains an unmet clinical need.

We describe a deep learning model for detecting elevated mPCWP from the standard 12-lead ECG alone. We also derive a quantitative metric that indicates when the model's prediction is most untrustworthy, allowing health care providers to identify when the model is likely to yield a misleading result. The resulting unreliability score is essential because deep neural networks are generally opaque to mechanistic explanation, so it not possible to validate a model output by examining its internal logic, unlike for a mechanistic model or a human decision-maker.

METHODS

DATA ACQUISITION. Our primary data set consisted of 6,739 right heart catheterization procedures from 4,304 patients, occurring between January 2010 and October 2020: our right heart catheterization (RHC) data set. These data correspond to an in-house registry at the Massachusetts General Hospital (MGH), where pressures were documented at the time of the catheterization procedure. This population only includes procedures for which same-day ECGs were available. In addition, any ECGs containing nonphysical values (eg, voltage >5 mV in magnitude in any lead) were removed.

FIGURE 1 Training and Testing Procedure



Data sets used for training and evaluating model 1, model 2, and RHCNet. The model with the best performance on the test set is also evaluated using the holdout set. ECG = electrocardiogram; RHCNet = Right Heart Catheterization Network.

Standard pressures were measured in each procedure: right atrial pressure (RAP), right ventricular pressure (RVP), pulmonary arterial pressure (PAP), and PCWP. For each site, systolic and/or diastolic pressures were often recorded in addition to the mean pressure. Only the mean pressure was utilized in this study. Cardiac output (CO), measured via thermodilution, was also obtained. We only utilized data from patients who had mean PAP (mPAP), mPCWP, and CO measurements documented. Pulmonary vascular resistance (PVR) was computed for each set of measurements using the mPAP, mPCWP, and CO. We binarized pressure measurements using a cutoff of 20 mmHg for the mPAP and 15 mmHg for the mPCWP. We binarized the CO and PVR with cutoffs of 4 L/min and 3 Wood units, respectively. For each measurement, a label of *zero* indicates a value below the given threshold and a label of *one* indicates a value above the threshold. Most of the ECGs extracted were sampled at 500 Hz. Any ECGs sampled at 250 Hz were upsampled by a factor of 2 via linear interpolation. Each set of results from RHC was matched to the first ECG taken on the day of catheterization.

Initial results suggested that model performance could be improved by first pre-training a model using a larger data set of ECGs for whom RHC data were not available. We used a previously constructed registry of ECGs, also derived from the MGH. For each patient in this registry, the most recent ECG was selected, provided that it was sampled at 500 Hz. All ECGs within the registry had ancillary data, including age, sex, and interval durations that were derived from automatic machine reads of the ECG at the time of acquisition. ECGs in this cohort correspond to patients aged between 20 and 90 years. The final cohort consists of 242,216 ECGs, each from a unique patient. The average age in this population was 56 ± 17 years, and 52% of the population was male. We refer to this registry as our pre-training data set.

DEVELOPMENT OF ELEVATED mPCWP DETECTION MODELS. We divided the RHC data set into a development set (5,390 samples) and a holdout set (1,349 samples). The development set is split into training (4,304 samples), validation (546 samples), and testing sets (540 samples) as shown in Figure 1. We developed 3 models to infer mPCWP from ECG data and evaluated each on the test set. The holdout set is used to further evaluate the best performing model, Right Heart Catheterization Network (RHCNet) (Figure 1). All data sets are divided by patient to avoid overlapping patients between data sets, and the rate of

TABLE 1 Main Indications for Catheterization in the Development Data Set and in the Holdout Data Set

Main Indication for RHC	Number of Cases in the Development Set (%)	Number of Cases in the Holdout Set (%)
Heart failure	2,376 (44.1)	552 (40.9)
Transplant	1,604 (29.8)	427 (31.7)
Coronary artery disease	731 (13.6)	163 (12.1)
Valvular disease	318 (5.9)	60 (4.4)
Pulmonary hypertension	155 (2.9)	62 (4.6)
Other pulmonary disease	66 (1.2)	37 (2.7)
Pericardial disease	20 (0.4)	0 (0.0)
Congenital defect	7 (0.1)	13 (1.0)
Electrical dysfunction	7 (0.1)	3 (0.2)
Other	106 (1.9)	32 (2.4)

Values are n (%).
RHC = right heart catheterization.

elevated mPCWP is roughly equivalent in each data set. The training set is used to optimize the model parameters, the validation set ("Val" in Figure 1) is used to determine when the training should be terminated, and the test set is used to evaluate the trained model.

Three models were developed during the course of this work:

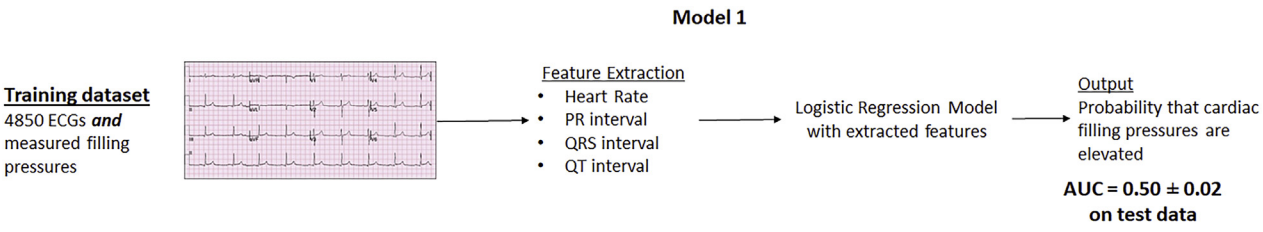
1. Model 1 is a logistic regression model. The inputs to this model are ECG interval data. These were extracted by the ECG acquisition systems-automated interval extraction algorithm. The model is trained to detect when the mPCWP >15 mmHg only using extracted intervals (ie, heart rate, PR interval, QRS interval, and QT interval).
2. Model 2 is a deep learning model (convolutional neural network, or CNN, see Supplemental Figure 1) that was trained for a multitask classification problem. The model tries to learn when each of 4 hemodynamic parameters are elevated (mPCWP >15 mmHg, mPAP >20 mmHg, PVR >3 Wood units, and CO >4 L/min). Although we are primarily interested in the mPCWP, the other estimated hemodynamic quantities are used to develop an unreliability score, which we discuss in subsequent sections.
3. RHCNet is also a deep learning model that was trained to accomplish the same multitask classification problem as model 2. RHCNet differs in that training was done in 2 stages: 1) A CNN is first trained to infer interval durations, including the RR interval, QRS interval, QT interval, and the PR interval, from the ECG alone; 2) In the second stage, the model (after pre-training) was modified by truncating it at its penultimate layer and then appending with additional dense layers. This new model was then trained again using the development data set (Figure 1).

To calculate statistical measures of uncertainty for all model results, ECGs were drawn from the development test set uniformly at random, with replacement, to produce 10 bootstrapped data sets of equal size to the original data. The same was done for the holdout data set. Error values of performance metrics represent 1 standard deviation computed across the results on the 10 bootstrapped data sets.

UNRELIABILITY SCORE. We developed an unreliability score that identifies subgroups where model predictions are poor. Essentially, the method computes the probability that the mPAP is elevated using 2 approaches. If the 2 approaches disagree for a specific ECG, then we propose that the model is unreliable for that ECG. Essentially, the method computes the probability that the mPCWP is elevated—our main prediction task—as well the probability that mPAP, PVR, and CO are abnormal. As all of these quantities are inter-related, the uncertainty score captures how consistent all of these measures are. If they are inconsistent, we postulate that the model prediction in question is untrustworthy. In practice, we compare the p(mPAP >20 mmHg), an explicit model output, to the probability that an ECG is associated with pulmonary hypertension. The probability of pulmonary hypertension, p(PH), accounts for patients with isolated pre-capillary PH, isolated post-capillary PH, and also mixed states where there is both pre- and post-capillary hypertension—hence the calculation of p(PH) requires knowledge that p(PVR >3 Woods Units) and p(mPCWP >15 mmHg). The associated probabilistic formalism is shown in the Supplemental Information. The unreliability score, U(ECG), is therefore a function of the difference between p(mPAP >20 mmHg) and p(PH). We note that $0 \leq U(ECG) \leq 1$, where higher values indicate more untrustworthy predictions.

To evaluate the utility of the unreliability score, data were sorted by U(ECG). Model outputs were split into 2 groups based on the U(ECG) score: those in the highest decile of scores and all others. The group with the higher scores represents the most unreliable predictions. Area under the receiver operating characteristic (AUC) and Brier scores were computed for each of these 2 groups. The Brier score is the average error of the model—lower Brier scores correspond to

FIGURE 2 ECG Interval Logistic Regression Model



In model 1, ECG intervals, derived from the machine read of each ECG, are used to train a logistic regression model to predict the probability that cardiac filling pressures are elevated. This model achieves an AUC of 0.50 ± 0.02. AUC = area under the receiver operating characteristic curve; ECG = echocardiogram.

lower, on average, errors. These metrics were computed for each bootstrap, and the standard deviation was computed for each metric from those results.

MODEL EVALUATION. The model with the highest discriminatory ability on the test data was also evaluated on the holdout data set of 1,349 clinical encounters (Figure 1). All statistical measures on the holdout set were computed in the same way as for the development test set, with 10 bootstraps generated by drawing uniformly at random, with replacement, from the holdout data set.

STATISTICAL ANALYSIS. Area under the receiver operating curve (AUC) is computed as follows: a sample $x_i^{(+)}$ is drawn uniformly at random from the positive-labeled data, and a sample $x_j^{(-)}$ is drawn from the negative-labeled data. Then, for the model f, which outputs a probability between 0 and 1, the following expression is evaluated:

$$f\left(x_i^{(+)}\right) > f\left(x_j^{(-)}\right)$$

This is repeated for N pairs of samples. The AUC score corresponds to the percentage of pairs of samples for which the aforementioned expression holds, that is,

$$AUC = \frac{1}{N} \sum_{\left\{x_i^{(+)}, x_i^{(-)}\right\}} 1\left[f\left(x_i^{(+)}\right) > f\left(x_j^{(-)}\right)\right]$$

where 1[…] is the indicator function. The AUC reflects the discriminatory ability of the model.[22]

Sensitivity values were calculated as the true positive rate—the rate of correct identification of positive cases. Sensitivities were computed using different model thresholds, and the corresponding specificities (true negative rates) were computed for each threshold. For a given sensitivity, specificity, and prevalence of elevated mPCWP, the positive predictive value (PPV) and negative predictive value (NPV) are computed as follows:

$$PPV = \frac{Sensitivity \cdot prevalence}{Sensitivity \cdot prevalence + (1 - specificity)(1 - prevalence)}$$

$$NPV = \frac{Specificity \cdot (1 - prevalence)}{Specificity \cdot (1 - prevalence) + (1 - sensitivity)prevalence}$$

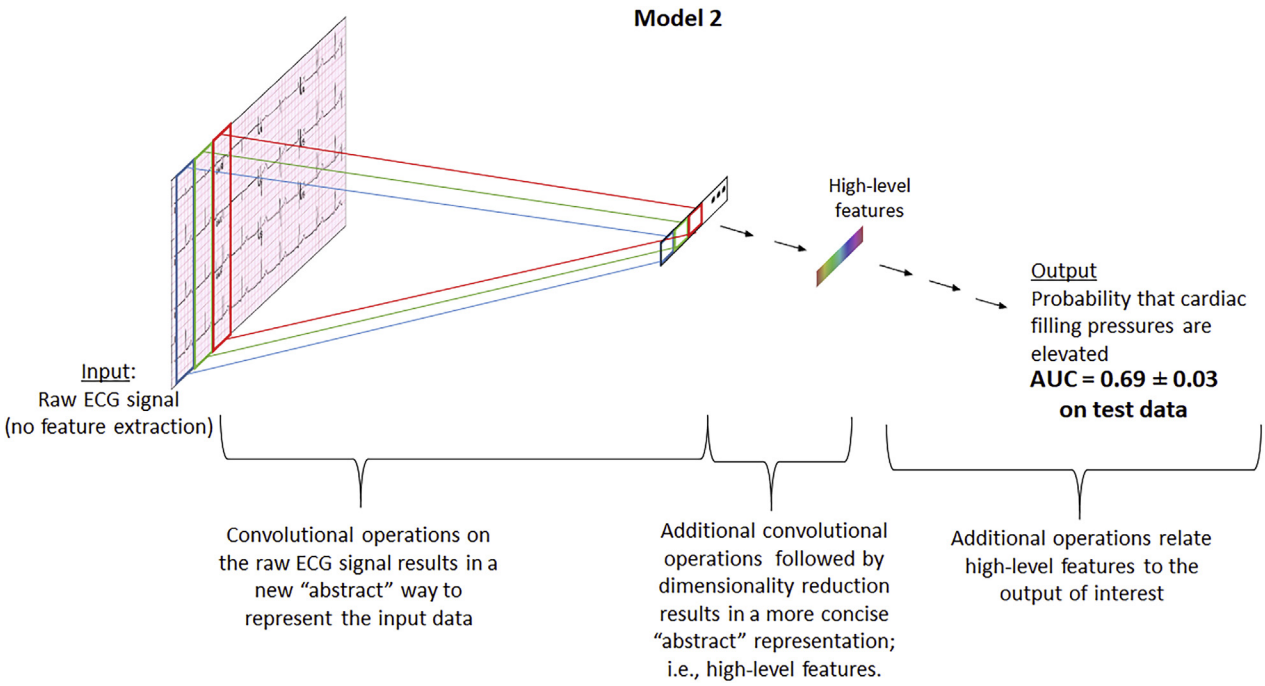
An unpaired, 2-tailed Student's *t*-test is used to compare results across bootstraps between 2 cohorts of interest, whenever variances between the 2 groups were similar. Otherwise, Welch's t-test was used.[23]

RESULTS

STUDY POPULATIONS EVALUATING MODEL DISCRIMINATORY ABILITY. Our development data set consists of patients who were referred for an RHC at the MGH (5,390 procedures from 3,446 patients) (Figure 1). In this cohort, 48% of procedures found an mPCWP >15 mmHg, where 15 mmHg is the upper limit of normal.[24] The average age at catheterization was 64 ± 4 years. Patient age ranged from 18 to 99 years. Sixty-four percent of the population was also male. Indications for catheterization are listed in Table 1. The most common indication was HF (44.1%), followed by heart transplant (29.8%), including surveillance endomyocardial biopsy and RHC in patients who are post-cardiac transplant.

We used these data to train 3 models to predict hemodynamic quantities. The first was a logistic

FIGURE 3 Randomly Initialized Deep Neural Network



In model 2, a deep neural network is trained to detect elevated mPCWP from the 12-lead ECG signal. Initial model parameters are set to random values and then modified to maximize model performance. Convolutional layers use specialized functions to extract information about the shape of the ECG. This yields an AUC of 0.69 ± 0.03. AUC = area under the receiver operating characteristic curve; ECG = echocardiogram; mPCWP = mean pulmonary capillary wedge pressure.

regression model that takes extracted ECG interval duration (PR, QRS, and QT intervals) and heart rate as input to predict when the mPCWP >15 mmHg (**Figure 2**). The resulting AUC was 0.50 ± 0.02, suggesting that these ECG features do not yield a model with discriminatory ability (**Figure 2**). We then trained a convolutional neural network—a type of deep learning model—to detect abnormal mPCWP, mPAP, CO, and PVR, using the raw ECG samples. The resulting AUC for predicting an elevated mPCWP was 0.69 ± 0.03 (**Figure 3**).

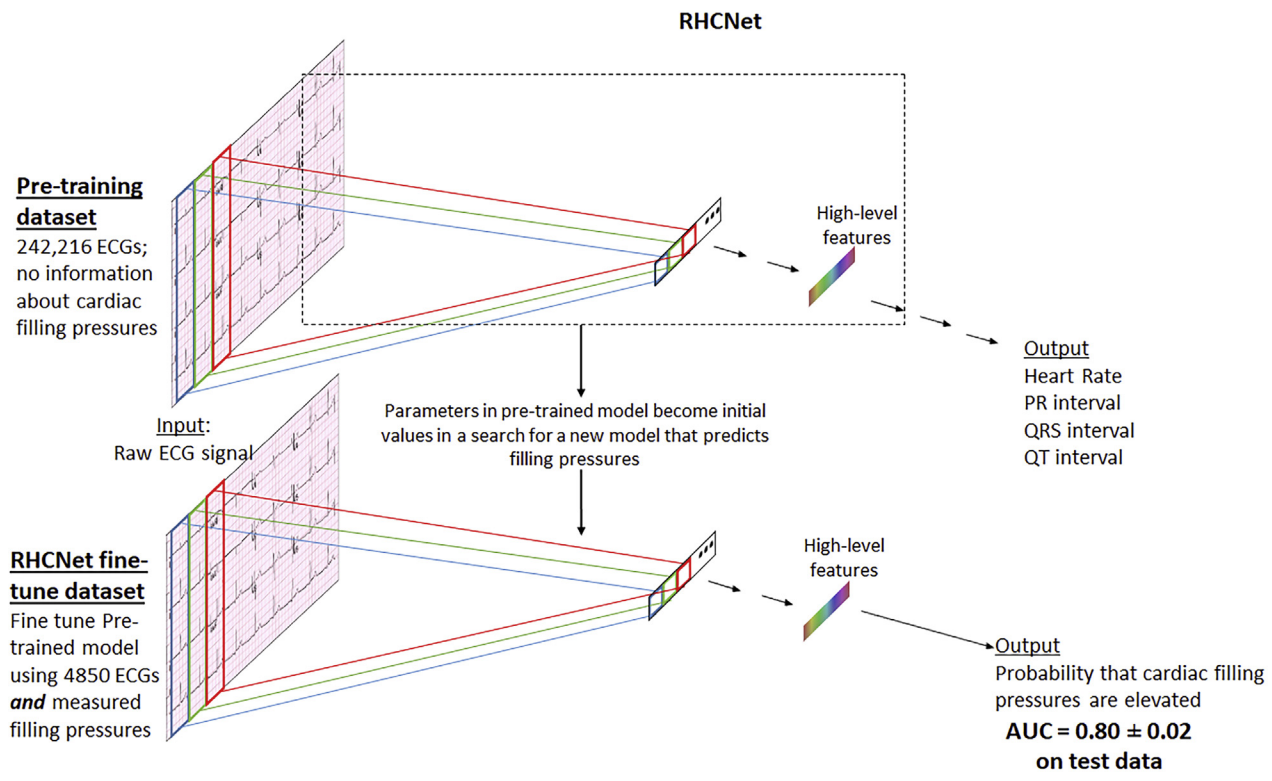
To further boost our performance, we leveraged a pre-training data set consisting of 242,216 ECGs derived from an in-house registry of patients at the MGH. The overall training strategy for this model, which we call RHCNet, is shown in **Figure 4**. The first "pre-training step" involves training the model to estimate the duration of ECG intervals (**Figure 4**, Supplemental Figure 2). Although extracted Intervals are not themselves predictive of abnormal mPCWPs (as demonstrated in model 1 results), we hypothesized that this pre-training step would yield "high-level" features that could be used to build a predictive model with good discriminatory ability (**Figure 4**). These high-level features represent complex functions of the original 12-lead ECG and therefore include more information than just the interval lengths themselves. The resulting model achieves an AUC for predicting an elevated mPCWP of 0.80 ± 0.02 on the test set data.

We evaluated the final model on the holdout data set of 1,349 samples, which had an average age at catheterization of 61 ± 16 years, and 66% were male. The proportion of procedures from patients with HF was 40.9%, and 31.7% of procedures were from transplant patients. In the holdout set, the AUC for inferring elevated mPCWP was 0.79 ± 0.01. ROC curves for all model outputs are shown in the Supplemental Figures 3 and 4.

EVALUATING RHCNet PREDICTIVE PERFORMANCE.

We evaluated model performance on patient subgroups of interest. We focused on the 2 indications in which we had more than 1,000 cases in the development set, to help ensure that we had enough data to compute robust metrics of performance, and used a

FIGURE 4 RHCNet Pre-training, Training, and Performance on Test Data



RHCNet is trained in 2 steps. First, a deep neural network model is trained to estimate ECG intervals from a large set of ECGs for which information about cardiac filling pressures was not available. Fine-tuning involves using the parameters from this model as initial values for additional training, using an independent data set of 12-lead ECGs that are matched to hemodynamic measurements. This yields an AUC of 0.80 ± 0.02 on the test set. Pre-training the model for a different task using a large data set results in a set of initial model parameters that give improved results. AUC = area under the receiver operating characteristic curve; ECG = echocardiogram; RHCNet = Right Heart Catheterization Network.

decision threshold corresponding to a sensitivity of 80%. In the holdout set, the PPV was 0.76 ± 0.02 in samples from patients with HF (prevalence: 61%) and the NPV was 0.92 ± 0.03 in samples from transplant patients (prevalence: 15%) (Table 2).

As the overall prevalence of an elevated mPCWP is high in our entire cohort, we explored how the model would perform in cohorts where the prevalence of disease is lower. Toward this end, we calculated both PPVs and NPVs for RHCNet using different values for the underlying prevalence. Since the PPV and the NPV are functions of the model's sensitivity and specificity, we first computed sensitivity and specificity values for the model using different decision thresholds (Figure 5A). PPVs and NPVs for prevalence values of 10, 20, 30, 40, and 50% are shown in Figures 5B and 5C. As expected, the PPV improves as the prevalence increases and the NPV increases when the prevalence decreases. In particular, when the sensitivity is 0.8 and the prevalence is 10%, the NPV is >0.97.

TABLE 2 RHCNet's Performance for Detecting mPCWP >15 mmHg Within Specified Cohorts of Interest

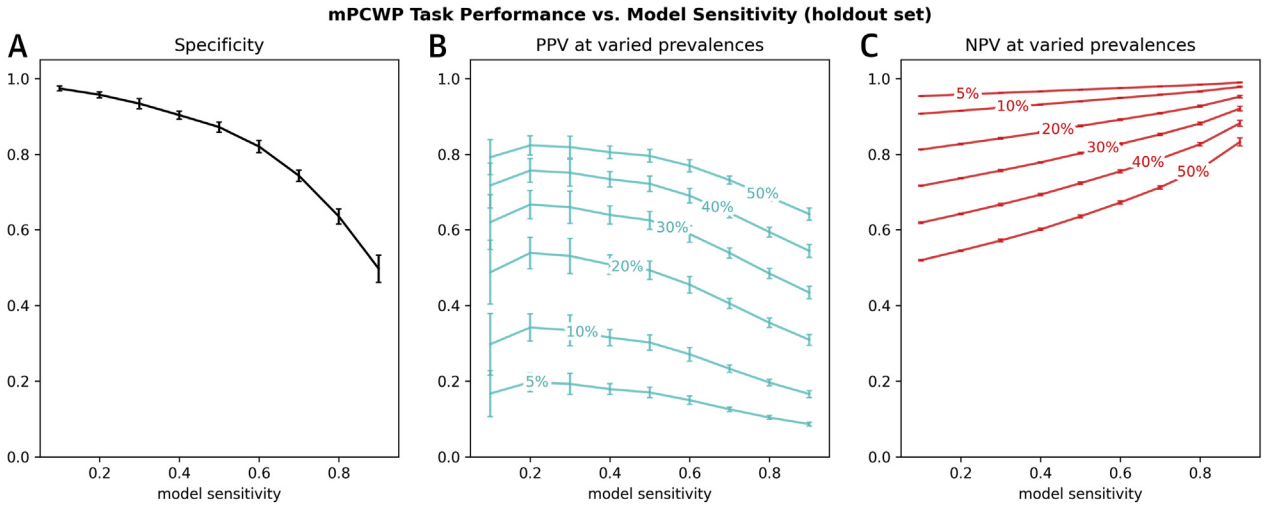
Data Set	Indication	PPV[a]	NPV[a]
Holdout set	Heart failure	0.76 ± 0.02	0.42 ± 0.04
	Transplant	0.19 ± 0.03	0.92 ± 0.03

Values are mean ± SD. [a]Using cutoff that achieves sensitivity of 0.80 for each bootstrap on the full test set.

mPCWP = mean pulmonary capillary wedge pressure; NPV = negative predictive value; PPV = positive predictive value; RHCNet = Right Heart Catheterization Network.

IDENTIFYING TRUSTWORTHY MODEL PREDICTIONS. We developed a score, $0 \leq U(ECG) \leq 1$, to identify subgroups associated with poor model performance. Large values of U are associated with relatively large model errors and therefore indicate an untrustworthy inference. We used the average model error (the Brier score) to quantify model performance in both reliable

FIGURE 5 RHCNet Performance as a Function of Sensitivity and Prevalence



(A) Specificity as a function of sensitivity; **(B)** PPV as a function of sensitivity (x-axis) and prevalence (5%, 10%, 20%, 30%, 40% and 50%); **(C)** NPV as a function of sensitivity and prevalence. NPV = negative predictive value; PPV = positive predictive value; RHCNet = Right Heart Catheterization Network.

and unreliable subgroups. In the holdout set, the average model error was high when the unreliability score was high (**Figure 6A**). Similarly, the discriminatory ability of the model was reduced when predictions have higher unreliability scores (**Figure 6B**).

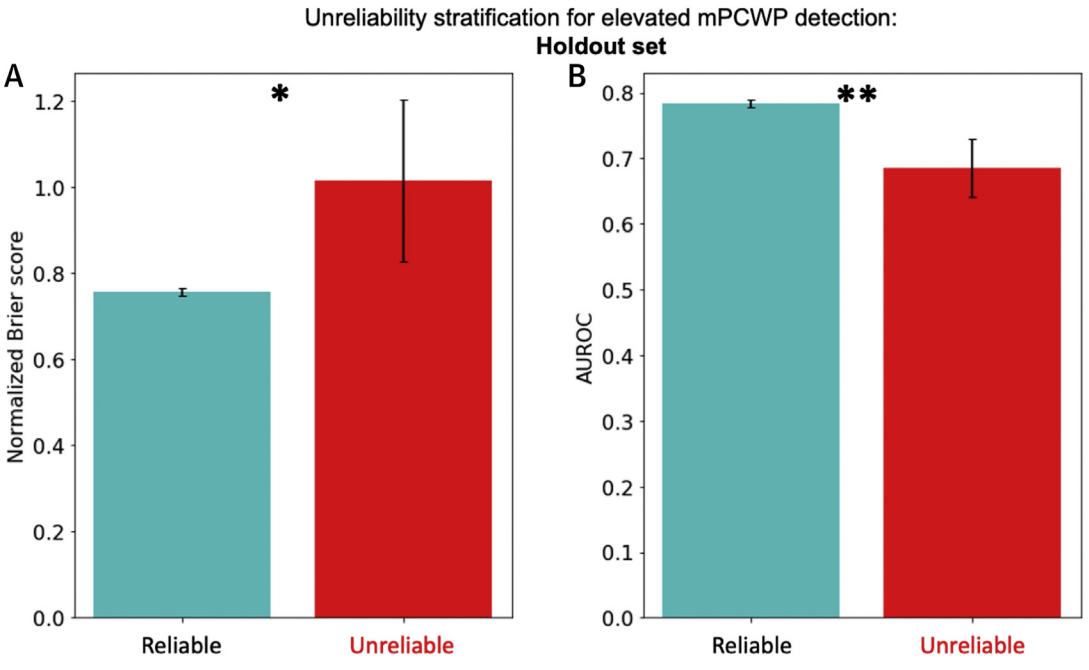
DISCUSSION

In this work we describe a novel deep learning model for detecting elevated mPCWP using the standard 12-lead ECG. To our knowledge, this is the first study to leverage deep learning to detect elevated mPCWP from the ECG alone. While the overall discriminatory ability of RHCNet is good (AUC: 0.80 on the development test set and 0.79 on the holdout set), the predictive performance of the model varies with indication. For patients with HF, where the prevalence of an elevated mPCWP is large, positive predictions are much more informative than negative predictions, while the opposite is true for post-cardiac transplant patients, where the prevalence of an elevated mPCWP is much smaller. To gauge model performance over the population at large, knowledge of the underlying prevalence is needed. Although estimates of the wider prevalence of elevated mPCWPs are lacking in the literature, we note that the prevalence of HF diagnoses in adults older than 60 years varies between 5 and 12%.[25] Assuming a prevalence of 5 to 10%, an RHCNet decision threshold yielding a sensitivity of 0.8 would achieve an NPV between 0.97 and 0.98 in this cohort, while still capturing 80% of the true positives.

Deep learning provides a platform to develop models that can utilize information in large and complex data sets to yield clinically useful insights, but such approaches are largely "black box methods." It is often challenging for even a computer scientist to understand precisely what a deep learning model has learned or why it arrives at a particular prediction. This remains an obstacle that limits broad acceptance by the clinical community.[26] As a first step toward understanding what RHCNet has learned, we use saliency maps—a method that discovers what portions of the input data a model tends to focus on when making a decision.[27,28] Calculated saliency maps suggest that our model tends to focus on the ventricular diastolic phase of the cardiac cycle—a finding in line with the notion that the mPCWP is often a good estimate of the left ventricular end-diastolic pressure[29] (Supplemental Figure 5). However, saliency maps, while useful, tell us where a model is looking, but do not fully explain how a model works. In short, no clear standard exists for the development and evaluation of explainable artificial intelligence systems.[30,31]

At their core, explainable models are attractive because they inspire trust—predictions that are achieved via a set of understandable and cogent steps are intuitively more trustworthy. In this vein, we developed a method that not only produces a prediction

FIGURE 6 **RHCNet Performance as a Function of Model Unreliability**



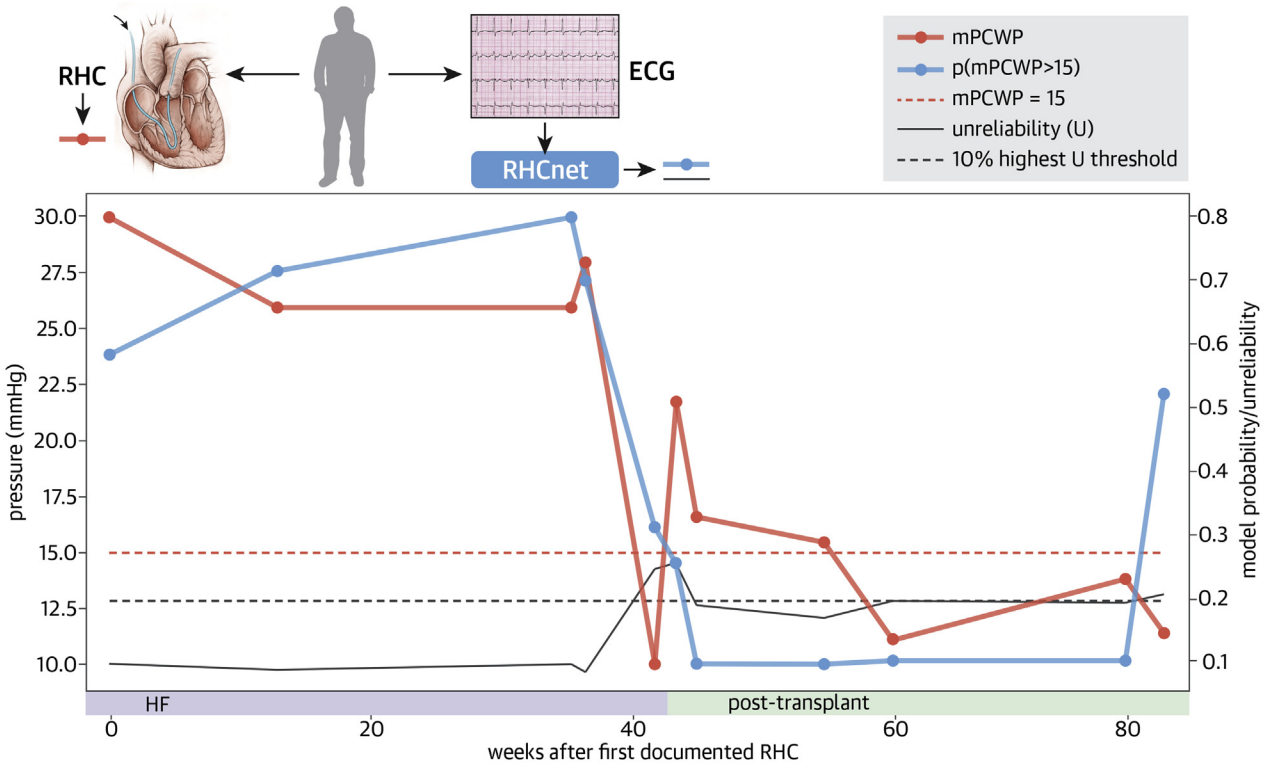
(A) Normalized Brier score (average model error) for inferring an elevated mPCWP in patients with high and low unreliability in the development set **(B)** and AUC for the same. The unreliable subgroup is defined as the set of all predictions having unreliable scores within the top decile. For simplicity, we refer to all other predictions that are relatively reliable. *$P < 10^{-2}$, **$P < 10^{-3}$. AUROC = area under the receiver operating characteristic curve; mPCWP = mean pulmonary capillary wedge pressure; RHCNet = Right Heart Catheterization Network.

but also aspires to identify when the model produces untrustworthy predictions. We do so by calculating an unreliability score that identifies when model predictions fall into a subgroup where model performance is likely to be poor. In addition to learning when the mPCWP is elevated, the model also learns information about other hemodynamic variables that are related to the mPCWP. If the model correctly captures the underlying physiology, the quantities estimated by the model should be internally consistent. When the model's predictions are incompatible, we assume that it has not correctly learned the underlying physiology for the given input, indicating that the corresponding predictions are untrustworthy. Using this principle as a guide, we derived a metric, *U*, that quantifies how internally consistent the estimated model predictions are. Our work demonstrates that while the overall discriminatory power of the model is good, model performance is poor in subgroups enriched with predictions that have high *U* values.

To illustrate how these data could be used in practice, we show the hemodynamic trajectory of an example patient from the holdout data set, alongside the model probability of elevated mPCWP and the unreliability metric (**Central Illustration**). For this patient, prior to orthotopic heart transplant, the measured mean wedge pressures are elevated, and model predictions agree with these observations. After transplantation, the patient's mPCWP drops and the inferred model probability decreases as well. Moreover, the most accurate predictions are associated with low unreliability scores. Together, the mPCWP and the associated unreliability score provide complementary information that may support the identification of when the mPCWP is truly elevated.

STUDY LIMITATIONS. Our model was developed on patients arising from a single tertiary care center; hence, additional studies are needed to determine its applicability to other settings. We have therefore made the model publicly available to facilitate its testing in different clinical environments (https://github.com/daphneschles/RHCnet). In addition, while we ensured that the time between 12-lead ECG acquisition and the cardiac catheterization was <24 hours, the precise time of catheterization is not known, introducing additional noise to the data due to hemodynamic changes that may have occurred

CENTRAL ILLUSTRATION **Evaluating Trends in mPCWP on a Per-Patient Basis**



Schlesinger DE, et al. JACC Adv. 2022;1(1):100003.

RHCNet predictions on serial ECGs for a patient in our holdout set. This patient had HF and then underwent orthotopic cardiac transplantation, as noted in the colored bar at the bottom of the figure. The x-axis represents the time from the patient's first RHC. At each time point a 12-lead ECG was obtained and RHCNet was used to estimate the probability that the mPCWP was elevated (blue), as well as the unreliability of that prediction (black), using only the 12-lead ECG. The right-hand axis corresponds to the model output and unreliability, both inferred from a 12-lead ECG. The left-hand axis corresponds to the measured mPCWP, acquired via right heart catherization on the same day that the ECG was acquired (red line). Overall, RHCNet tracks the patient's mPCWP and subsequent improvement after transplant. Moreover, high unreliability scores (near the threshold corresponding to the 10% highest unreliability) are more likely to be incorrect. These data suggest that RHCNet may enable serial non-invasive tracking of central hemodynamics using the 12-lead ECG. ECG = electrocardiogram; HF = heart failure; mPCWP = mean pulmonary capillary wedge pressure; PAC = pulmonary artery catheter; RHC = right heart catheterization; RHCNet = Right Heart Catheterization Network.

between the acquisition of the ECG and the catheterization procedure. We expect that better results could be obtained if each 12-lead ECG were recorded just prior to catheterization. Our model generates a binary classification of the mPCWP based on a clinically important threshold rather than predicting the precise value of the mPCWP itself. Larger data sets, and prospective studies, are required to determine the predictive value of deep learning models for estimating advanced hemodynamic parameters.

CONCLUSIONS

Our results suggest that a machine learning model is able to identify when the mPCWP is elevated using information from the ECG alone. The model has the potential to be an effective screening tool for the detection of elevated left-sided filling pressures in selected patients. Further studies are needed to establish the potential clinical applications of the method in real-world applications.

ACKNOWLEDGMENTS All retrospective data analyses for this study were approved by the Institutional Review Board (IRB) at the MGH (protocol #2020P000132). Data sharing: Code, model weights, and example input are available at https://github.com/daphneschles/RHCnet.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

This work was supported by a grant from Quanta Computers Inc. to support the development of machine learning methods that can

improve the care of patients, with cardiovascular disease, from diverse populations. Quanta computers Inc. played no role in designing the experiments, analyzing the results, or writing/reviewing the manuscript. The authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr Collin M. Stultz, Massachusetts Institute of Technology & Massachusetts General Hospital, MIT, Building 36-796, 77 Massachusetts Ave, Cambridge, Massachusetts 02139, USA. E-mail: cmstultz@mit.edu.

PERSPECTIVES

COMPETENCY IN SYSTEMS-BASED PRACTICE: Deep learning applied to the 12-lead surface ECG can be used to estimate when the mPCWP is elevated in patients with suspected cardiac disease.

TRANSLATIONAL OUTLOOK: Further prospective trials of RHCNet are needed to determine the clinically utility of using deep learning to estimate central pressures in patients with cardiac disease.

REFERENCES

1. Drazner MH, Hellkamp AS, Leier CV, et al. Value of clinician assessment of hemodynamics in advanced heart failure: the ESCAPE trial. *Circ Heart Fail*. 2008;1:170–177.

2. Nair R, Lamaa N. *Pulmonary Capillary Wedge Pressure*. Treasure Island (FL): StatPearls; 2020.

3. Binanay C, Califf RM, Hasselblad V, et al. Evaluation study of congestive heart failure and pulmonary artery catheterization effectiveness: the ESCAPE trial. *JAMA*. 2005;294:1625–1633.

4. Shah MR, Hasselblad V, Stevenson LW, et al. Impact of the pulmonary artery catheter in critically ill patients: meta-analysis of randomized clinical trials. *JAMA*. 2005;294:1664–1670.

5. Sandham JD, Hull RD, Brant RF, et al. A randomized, controlled trial of the use of pulmonary-artery catheters in high-risk surgical patients. *N Engl J Med*. 2003;348:5–14.

6. Saxena A, Garan AR, Kapur Navin K, et al. Value of hemodynamic monitoring in patients with cardiogenic shock undergoing mechanical circulatory support. *Circulation*. 2020;141:1184–1197.

7. Guyton AC, Lindsey AW. Effect of elevated left atrial pressure and decreased plasma protein concentration on the development of pulmonary edema. *Circ Res*. 1959;7:649–657.

8. Cooper LB, Mentz RJ, Stevens SR, et al. Hemodynamic predictors of heart failure morbidity and mortality: fluid or flow? *J Card Fail*. 2016;22:182–189.

9. Drake RE, Doursout MF. Pulmonary edema and elevated left atrial pressure: four hours and beyond. *Physiology*. 2002;17:223–226.

10. Yu C-M, Wang L, Chau E, et al. Intrathoracic impedance monitoring in patients with heart failure. *Circulation*. 2005;112:841–848.

11. Sprung CL, Pozen RG, Rozanski JJ, Pinero JR, Eisler BR, Castellanos A. Advanced ventricular arrhythmias during bedside pulmonary artery catheterization. *Am J Med*. 1982;72:203–208.

12. Kearney TJ, Shabot MM. Pulmonary artery rupture associated with the Swan-Ganz catheter. *Chest*. 1995;108:1349–1352.

13. Karmali KN, Goff DC Jr, Ning H, Lloyd-Jones DM. A systematic examination of the 2013 ACC/AHA pooled cohort risk assessment tool for atherosclerotic cardiovascular disease. *J Am Coll Cardiol*. 2014;64:959–968.

14. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18:500–510.

15. Tison GH, Zhang J, Delling FN, Deo RC. Automated and interpretable patient ECG profiles for disease detection, tracking, and discovery. *Circ Cardiovasc Qual Outcomes*. 2019;12:e005289.

16. Porumb M, Stranges S, Pescapè A, Pecchia L. Precision medicine and artificial intelligence: a pilot study on deep learning for hypoglycemic events detection based on ECG. *Sci Rep*. 2020;10:170.

17. Kwon J-M, Cho Y, Jeon K-H, et al. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. *Lancet Digital Health*. 2020;2:e358–e367.

18. Kwon J-M, Kim K-H, Medina-Inojosa J, Jeon K-H, Park J, Oh B-H. Artificial intelligence for early prediction of pulmonary hypertension using electrocardiography. *J Heart Lung Transplant*. 2020;39:805–814.

19. Kusunose K, Hirata Y, Tsuji T, Kotoku JI, Sata M. Deep learning to predict elevated pulmonary artery pressure in patients with suspected pulmonary hypertension using standard chest X ray. *Sci Rep*. 2020;10:19311.

20. Hellenkamp K, Unsöld B, Mushemi-Blake S, et al. Echocardiographic estimation of mean pulmonary artery pressure: a comparison of different approaches to assign the likelihood of pulmonary hypertension. *J Am Soc Echocardiogr*. 2018;31:89–98.

21. Nagueh SF, Smiseth OA, Appleton CP, et al. Recommendations for the evaluation of left ventricular diastolic function by echocardiography: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *J Am Soc Echocardiogr*. 2016;29:277–314.

22. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Lett*. 2006;27:861–874.

23. Welch BL. The generalization of 'student's problem when several different population variances are involved. *Biometrika*. 1947;34:28–35.

24. Hurst JW, Rackley CE, Sonnenblick EH, Wenger NK. *The Heart: Arteries and Veins*. New York: McGraw-Hill, Inc; 1990.

25. Roger VL, Go AS, Lloyd-Jones DM, et al. Heart disease and stroke statistics–2012 update: a report from the American Heart Association. *Circulation*. 2012;125:e2–e220.

26. Schlesinger DE, Stultz CM. Deep learning for cardiovascular risk stratification. *Curr Treat Options Cardiovasc Med*. 2020;22:15.

27. Kadir T, Brady M. Saliency, scale and image description. *Int J Computer Vis*. 2001;45:83–105.

28. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Workshop at *International Conference on Learning Representations*, 2014.

29. O'Quin R, Marini JJ. Pulmonary artery occlusion pressure: clinical physiology, measurement, and interpretation. *Am Rev Respir Dis*. 1983;128:319–326.

30. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, The Precise Qc. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20:310.

31. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Machine Intelligence*. 2019;1:206–215.

KEY WORDS deep learning, ECG, pulmonary artery occlusion pressure, pulmonary artery wedge pressure, pulmonary capillary wedge pressure

APPENDIX For a supplemental appendix and figures, please see the online version of this paper.