Artificial intelligence for hemodynamic monitoring with a wearable electrocardiogram monitor

Check for updates

Daphne E. Schlesinger[1,2,3,4,10], Ridwan Alam [2,3,4,10], Roey Ringel [4,5], Eugene Pomerantsev[4,6], Srikanth Devireddy[6,7], Pinak Shah[6,7], Joseph Garasic[6,7] & Collin M. Stultz [1,2,3,4,6,8,9] ✉

Abstract

Background The ability to non-invasively measure left atrial pressure would facilitate the identification of patients at risk of pulmonary congestion and guide proactive heart failure care. Wearable cardiac monitors, which record single-lead electrocardiogram data, provide information that can be leveraged to infer left atrial pressures.
Methods We developed a deep neural network using single-lead electrocardiogram data to determine when the left atrial pressure is elevated. The model was developed and internally evaluated using a cohort of 6739 samples from the Massachusetts General Hospital (MGH) and externally validated on a cohort of 4620 samples from a second institution. We then evaluated model on patch-monitor electrocardiographic data on a small prospective cohort.
Results The model achieves an area under the receiver operating characteristic curve of 0.80 for detecting elevated left atrial pressures on an internal holdout dataset from MGH and 0.76 on an external validation set from a second institution. A further prospective dataset was obtained using single-lead electrocardiogram data with a patch-monitor from patients who underwent right heart catheterization at MGH. Evaluation of the model on this dataset yielded an area under the receiver operating characteristic curve of 0.875 for identifying elevated left atrial pressures for electrocardiogram signals acquired close to the time of the right heart catheterization procedure.
Conclusions These results demonstrate the utility and the potential of ambulatory cardiac hemodynamic monitoring with electrocardiogram patch-monitors.

Plain language summary

Heart failure is a common disorder that is challenging to manage. Appearance of symptoms can be subtle and dangerous and there are few tools for clinicians to estimate when a patient is likely to experience an episode of heart failure. Current methods to detect elevated pressure in the heart (one sign of oncoming failure) are invasive and can only be performed in an inpatient setting. A non-invasive, quick method for detecting higher heart pressure would be helpful for identifying worsening heart failure in the home environment. For this reason, we developed a computer method to detect elevated pressures inside the heart using a non-invasive signal from a wearable patch monitor device, the electrocardiogram (ECG, or EKG). Our results show our method provides a reliable, non-invasive way to measure heart pressures using data that can be obtained in the outpatient setting.

Heart failure (HF) is associated with a range of adverse outcomes, resulting in increased healthcare expenditures, morbidity, and mortality[1,2]. Although progress has been made with respect to the management of patients with HF, critical challenges remain. Over 30 percent of HF patients admitted to the hospital are readmitted within 90 days of discharge. Many of these readmissions are avoidable[3,4].

Patients who have symptoms of congestive HF present with elevated left atrial pressure (LAP), requiring medical therapy to reduce these pressures and relieve pulmonary congestion. The gold standard method for measuring the LAP is right heart catheterization (RHC)—an invasive procedure that requires the placement of a catheter attached to a pressure transducer into the right heart and pulmonary arteries. During RHC, a branch of the pulmonary artery can be "wedged", forming a static column of blood between the catheter and the left atrium. This measurement is referred to as the mean pulmonary capillary wedge pressure (mPCWP) and is a reliable estimate of the LAP and left ventricular diastolic pressure in most patients[5,6].

Elevated mPCWP is an independent predictor of adverse outcomes in patients with HF, and lowering the mPCWP is an important intervention that can reduce the probability of readmission[7,8]. Furthermore, several

[1]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA. [2]Research Laboratory of Electronics, MIT, Cambridge, MA, USA. [3]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA. [4]Division of Cardiology, Massachusetts General Hospital, Boston, MA, USA. [5]Boston University School of Medicine, Boston, MA, USA. [6]Harvard Medical School, Boston, MA, USA. [7]Division of Cardiovascular Medicine, Brigham and Women's Hospital, Boston, MA, USA. [8]Institute for Medical Engineering and Science, MIT, Cambridge, MA, USA. [9]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA. [10]These authors contributed equally: Daphne E. Schlesinger, Ridwan Alam. ✉e-mail: cmstultz@mit.edu

studies suggest that the mPCWP begins to rise before the onset of symptoms in HF patients[9–11]. Hence, the ability to track the mPCWP in the outpatient setting would enable physicians to identify high-risk patients and proactively initiate medical therapy, thereby circumventing hospital admission. However, RHC cannot be routinely performed in the outpatient setting. Non-invasive methods for estimating mPCWP suffer from similar challenges that limit their application in an ambulatory setting. For example, while evidence of elevated mPCWP can be garnered from non-invasive cardiac Doppler ultrasound, a skilled sonographer is needed to obtain the required Doppler profiles[12,13]. In addition, accurate estimation of mPCWP from the physical exam alone is challenging and often unreliable, even when performed by seasoned experts[14]. Lastly, simple models that use clinical demographics, including heart rate parameters, perform poorly with respect to LAP estimation[15].

Due to these limitations, several devices have been developed to estimate mPCWP at home[9,10,16]. The CardioMEMS HF system, for example, is used to measure the diastolic pressure in the pulmonary artery, a surrogate for the LAP, and its use has been shown to reduce the hospitalization rate of patients with advanced HF by 50 percent[10]. Despite its benefits in terms of patient outcomes, using CardioMEMS requires an invasive procedure to implant the device within the main pulmonary artery, entailing some risk. The development of a reliable, easy-to-use non-invasive method for monitoring mPCWP, and cardiac hemodynamics more generally, would transform the management of patients with heart failure.

In recent years, artificial intelligence (AI) has shown promise in healthcare applications, including the analysis of medical time-series data to predict patient outcomes. The use of machine learning algorithms can facilitate remote patient monitoring in a range of cardiovascular diseases, via wearable devices measuring the single-lead electrocardiogram (ECG). Previous applications include the detection of atrial fibrillation[17–19], cardiac ischemia[20], long-QT syndrome[21], and reduced ejection fraction[22]. Many of these studies have focused on ECG signals that are acquired from smartwatch devices, which can yield noisy signals when the subject is not at rest[23]. Deep learning models trained on clinical ECG have been shown to effectively extrapolate to signals from insertable cardiac monitors, highlighting the potential for device-neutral solutions[24]. Wearable patch-monitors, which are applied to the chest wall, provide higher quality data, but the utility of ECG signals arising from these devices for machine learning tasks has not been widely explored.

Previously, we proposed a deep learning model that identifies when the mPCWP is elevated using the 12-lead ECG[15,25]. Although these studies demonstrate that the surface ECG has information for identifying abnormal cardiac hemodynamics, the 12-lead ECG is typically obtained during an office visit or in the inpatient setting, and requires the placement of 10 electrodes on the body. Consequently, it is not routinely used for outpatient monitoring. In this work, we propose a system to estimate mPCWP from Lead I of the ECG, called the Cardiac Hemodynamic AI monitoring System (CHAIS). CHAIS uses a deep neural network to analyze a single-lead ECG signal and infer if the patient's hemodynamics are abnormal. In the current study, we used a threshold of 18 mmHg as this value has been shown to be an independent predictor of adverse outcomes in patients with a prior diagnosis of HF[7,26,27]. We demonstrate the system's ability to predict an elevated mPCWP on retrospective datasets of patients from two large hospitals in Boston, MA. We also prospectively evaluate the model on a cohort of patients who wore a commercially available wearable patch-monitor prior to invasive RHC. Using the ground truth mPCWP data from the procedure, we directly evaluate how well the method performs when using data obtained from a wearable patch-monitor.

Methods

Retrospective data acquisition

This study was approved by the Massachusetts General Brigham Institutional Review Board (IRB protocol #2020P000132). The requirement to obtain informed patient consent was waived for this retrospective study. Data were collected for patients who underwent cardiac catheterization at

Table 1 | Dataset characteristics

Datasets	Development	Internal-holdout	External-validation	Wearable patch-monitor
No. of ECG	5390	1349	4620	83
No. of Patients	3446	858	2573	83
Age (mean ± STD)	64 ± 15	61 ± 16	57 ± 15	67 ± 13
Male (%)	63.9	65.5	64.6	67.5
Prevalence (%)	36.3	34.3	26.6	35.0

ECG electrocardiogram, *mPCWP* mean capillary wedge pressure, *Age* Age at catheterization, *STD* standard deviation, *Prevalence* Percent of patients with mPCWP > 18 mmHg.

two institutions. At Massachusetts General Hospital (MGH), data from RHC procedures performed between January 2010 and October 2020 were collected and matched to the first 12-lead ECG from the same calendar day as the procedure ($N = 6739$). Data from the Brigham and Women's Hospital (BWH) were collected between June 2009 and January 2019, and likewise matched to a same-day ECG ($N = 4620$). BWH patients who were also provided care at MGH were excluded from the BWH dataset to guarantee non-overlap between training and external-validation datasets. Demographic details for the patient populations are provided in Table 1. Samples were further categorized by indication for catheterization using the procedure described in Supplementary Methods (see Supplementary Table S1).

Data pre-processing

The ECG acquisition machines used at MGH and BWH acquire the ECG signals at either 500 or 250 Hz; we resample all signals recorded at 250 Hz using linear interpolation to 500 Hz. ECGs containing voltage values greater than 5 mV in magnitude were removed, as these likely represent artifacts. For model training, each ECG lead was standardized (Z-scored) using that lead's mean and variance.

Our original model focused on predicting several hemodynamic parameters from the 12-lead ECG. In this study, we focus on the performance of the model with respect to detecting a mPCWP greater than 18 mmHg.

Model pre-training

CHAIS is a single-lead adaptation of the 12-lead ECG residual neural network for estimating cardiac hemodynamics[15]. We initially pre-train the model to regress the PR, QRS, and QT intervals and the heart rate from the 12-lead ECG, using a cohort of 242,216 patients at MGH. The PR, QRS, and QT intervals for each ECG were measured by the ECG acquisition machines (GE and Philips) and reviewed by attending cardiologists. These features are stored with other clinical metadata in the dataset. There is no overlap between the patients in the pre-training cohort and those in the hemodynamics-matched MGH development and internal-holdout datasets. More details on the model parameters and the model architecture are provided in Supplementary Methods and Supplementary Fig. S1.

Fine-tuning on single-lead data

To fine-tune on single-lead data, Lead-I of the ECG is replicated 12 times to form a tensor that is 12×5000 samples (10 s at 500 Hz). Empirically, this procedure yielded better results relative to using a single 1×5000 tensor as input. The final 12×5000 tensor was Z-scored before training and testing.

For fine-tuning on the retrospective single-lead data, the pre-trained model, up to the 416-node layer, is appended with several dense layers (Fig. S1c), which are randomly initialized. The fine-tuning process involved learning weights for the end-to-end model, including the pre-trained layers. Additional training details are provided in Supplementary Methods.

The data from MGH were divided into a development set and an internal-holdout set, with 80 percent in the former (5390 samples) and 20 percent in the latter (1349 samples). The development set is split into training (80 percent, 4304 samples), validation (10 percent, 546 samples),

and internal-test (10 percent, 540 samples) sets. The whole BWH dataset is used solely for validation, and is referred to as the external-validation set (4620 samples). There are no overlapping patient data in any of these datasets (development, internal-holdout, and external-validation); i.e., data from a single patient only appears in one dataset.

Prospective data collection
This study was approved by the Mass General Brigham Institutional Review Board (IRB protocol #2016P001855). After obtaining patient consent (in accordance with our approved protocol) prospective data were collected from patients scheduled to undergo cardiac catheterization with IRB approval. Demographic details of these patients are provided in Table 1, and further categorization by indication for catheterization is presented in Supplementary Table S2. Our prospective study uses a commercially available ECG monitor (QOCA Portable ECG 101 Patch-monitoring Device). The monitor records single-lead ECG data in 12-bit resolution at a sampling rate of 256 Hz. Eligible inpatients at the MGH who were admitted to the cardiovascular step-down unit, and who were scheduled for a RHC, were consented by clinical research staff. The ECG patch-monitor device was placed below the left clavicle ~24 h before and removed the morning of their scheduled procedure. The device was oriented to achieve ECG signals similar to ECG Lead I. Once removed, information from the device was downloaded to an encrypted laptop and uploaded to a secure server for analysis.

Patch-monitor device data pre-processing
ECG signals from the patch-monitor device were resampled to 500 Hz, to match the sampling rate used in the 12-lead ECGs on which the model was originally trained. As CHAIS is trained on 10-s segments of ECG data, our goal was to identify 10 s of high-quality signal from the patch-monitor that could be used as input to the CHAIS algorithm. Towards this end, we segmented the ECG data arising from the patch-monitor into 5-min intervals and identified the 10-s within each interval that had the highest signal-to-quality index, which is calculated using the NeuroKit2 Python package[21]. If no 10-s segment within a given 5-min interval was above 0.5, then that 5-min interval was discarded. Additional details regarding the pre-processing and quality evaluation of the patch-monitor device data are provided in Supplementary Methods.

Zero-shot transfer learning to patch-monitor device data
We applied CHAIS to the patch-monitor device data with no fine-tuning. The patch-monitor device ECGs were pre-processed in the same manner as those for the retrospective studies: each high-quality 10-s ECG segment was normalized by its mean and variance in voltage. Probability values were inferred from the pre-processed samples. We used the highest quality 10-s ECG segment from each 5-min interval in the dataset, wherever a sufficiently high-quality segment was available, yielding approximately one CHAIS prediction for each 5-min interval.

Evaluation metrics
To assess how CHAIS could be used in practice, we computed the sensitivity and specificity using the internal and external datasets. To compute the sensitivity, one must first choose a cutoff for the model output that defines a positive prediction—i.e., when the mPCWP is predicted to be greater than 18 mmHg—and uses this value to compute the true positive rate (sensitivity) and the true negative rate (specificity). These cutoffs were derived from the combined development dataset, then applied to the internal-holdout dataset.

PPV and NPV can be computed for different population prevalence, or pre-test probability, values, using the following expressions[28]:

$$PPV = \frac{\text{sensitivity} \cdot \text{prevalence}}{\text{sensitivity} \cdot \text{prevalence} + (1 - \text{specificity}) \cdot (1 - \text{prevalence})} \tag{1}$$

$$NPV = \frac{\text{specificity} \cdot (1 - \text{prevalence})}{\text{specificity} \cdot (1 - \text{prevalence}) + (1 - \text{sensitivity}) \cdot \text{prevalence}} \tag{2}$$

We calculate the area under the receiver operating curve (AUROC) for presenting the model performance on the internal-holdout and the external-validation datasets. For the wearable ECG dataset from the prospective study, we calculate the AUROC as a function of time to the RHC procedure for respective patients. For each patient, using the method presented above (see subsection "Patch-monitor device data pre-processing"), we acquire one 10-s ECG signal for every 5 min over the whole duration of the hospital stay. From the absolute time of the RHC for that patient, we map each of those 10-s ECG to its corresponding relative time to RHC, referred here as the time-delta. This mapping allows us to look back at varying time-deltas from the RHC reference point. We look back at time-deltas ranging from 1 to 24 h prior to the RHC at an interval of 5 min. As shown in Supplementary Fig. S4, the number of patients who were monitored at a specific time before their RHC varies with the time-delta, hence the number of available ECG (N) varies with time-delta. For each of those time-deltas to RHC, we then calculate the AUROC over the available ECGs at that time-delta as presented in Fig. 4. Only time-deltas for which at least 10 samples were available were included. Error bars were computed as the standard error of the mean over 1000 stratified bootstraps for that data at each time-delta. Standard deviations for AUROC values and other statistics are computed over 1000 bootstraps, where the observed label prevalence is preserved within each bootstrap. Reported error bars in each table and figure correspond to the standard error of the mean.

Model trustworthiness score
To determine when the model is likely to yield a misleading result, we use the Shannon Entropy, in a manner similar to the method used in Raghu et al.[25]. Let $f_y(x)$ denote the model probability for one of the inference tasks, y, where x is a given input ECG. Then the entropy for a given prediction is:

$$H_y(x) = -f_y(x) \cdot \ln(f_y(x)) - (1 - f_y(x)) \cdot \ln(1 - f_y(x)) \tag{3}$$

This expression captures, in essence, how close the output model probability is to 0.5, with higher H_y reflecting a value closer to 0.5, and lower model trustworthiness. Note that $0 \leq H_y \leq \ln 2$. The entropy can be used to compute uncertainty for any of the four model targets, independently; we utilize this metric for characterizing the predictions of elevated wedge pressure (mPCWP). To threshold the uncertainty score, we use the value that splits the development dataset into the 90 percent lowest and 10 highest by uncertainty, so as to exclude only the most uncertain predictions.

Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results
CHAIS is trained to detect an elevated mPCWP from single-lead ECG data. Four datasets were used to develop and evaluate CHAIS. The development dataset, obtained from MGH, was used to train the model (Fig. 1a). An internal-holdout dataset, also obtained from MGH, and an external-validation dataset, obtained from the BWH, were used to evaluate the model (Fig. 1b). Finally, we prospectively collected single-lead ECG data using a commercially available ECG patch-monitor to further evaluate CHAIS performance on this ECG patch-monitor dataset (Fig. 1c). Characteristics of patients in each of these datasets are shown in Table 1, and the diagnostic breakdown of each dataset is shown in Supplementary Table S2.

Evaluation on internal-holdout set
CHAIS was trained using Lead I of the ECG, as this lead is commonly acquired by patch and wearable ECG monitoring devices[21,29–31]. We first
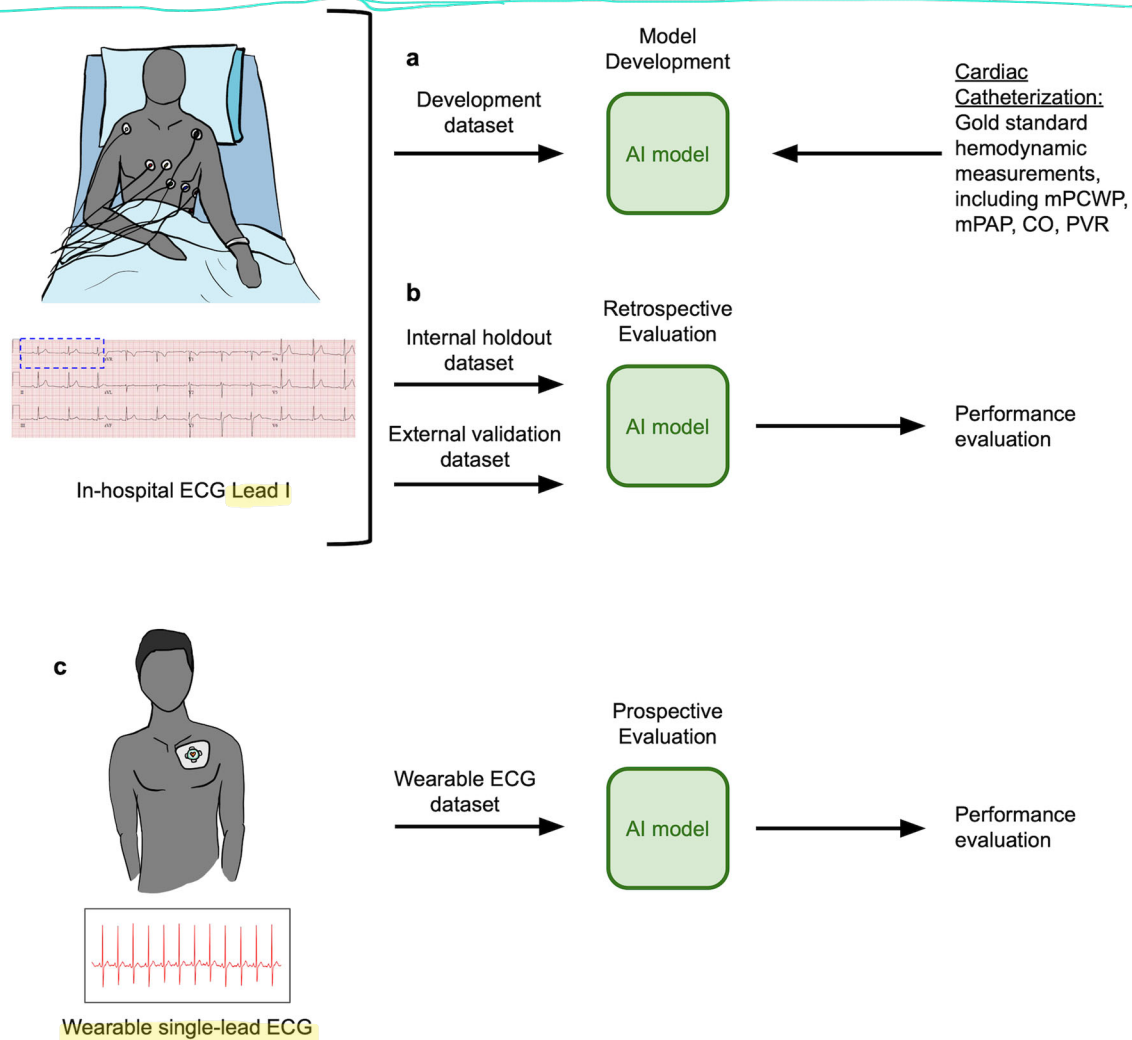
Fig. 1 | Model development and evaluation. a CHAIS is trained on an internal MGH development dataset, using 10 s, single-lead recordings (Lead I) extracted from the 12-lead electrocardiogram from patients with known cardiac hemodynamic measurements; **b** CHAIS is evaluated on an internal-holdout dataset and an external-validation dataset containing patients with ECG data (only Lead I is used) and known cardiac hemodynamics; **c** the model is then prospectively evaluated using ECG data acquired from patients who wore a commercially available ECG patch-monitor.

evaluated CHAIS using an internal-holdout dataset, which does not contain any patients from the development set. For detecting elevated mPCWP, the model achieved an AUROC of 0.80 on the internal-holdout dataset (Table 2).

As no model is perfect, it is important, when possible, to identify predictions that are likely to be incorrect. This can help practitioners determine when to trust a model output. Hence, as a secondary analysis, we

calculate the Shannon entropy using the CHAIS output to determine when the model is likely to yield a misleading result, in a manner similar to the method used in Raghu et al. [25]. We hypothesized that predictions associated with high entropy are more untrustworthy relative to predictions with low entropy. We define trustworthy predictions as those that have low entropies, where the entropy threshold is derived from the development dataset, as outlined in the section "Methods". The derived threshold was 0.6913123. This post-facto analysis of the model predictions can distinguish trustworthy predictions against those that are not sufficiently reliable. We observe, as shown in Table 2, that trustworthy predictions correspond to a subset where the model has higher discriminatory ability relative to untrustworthy predictions. For the task of identifying an elevated mPCWP in the internal-holdout set, the AUROC computed for the more trustworthy predictions was also 0.80 (the same as the performance on the whole internal-holdout set), whereas the AUROC is 0.52 for untrustworthy predictions. Such a trust score can be helpful for physicians in the decision-making process.

We also evaluated the model's predictive performance. Using a cutoff that achieves a sensitivity of 70 percent in the development dataset, we find a sensitivity of 71 percent, the associated specificity is 75 percent. The associated confusion matrix is presented in the Supplementary Information in Table S3. With a cutoff that achieves a sensitivity of 80 percent in the
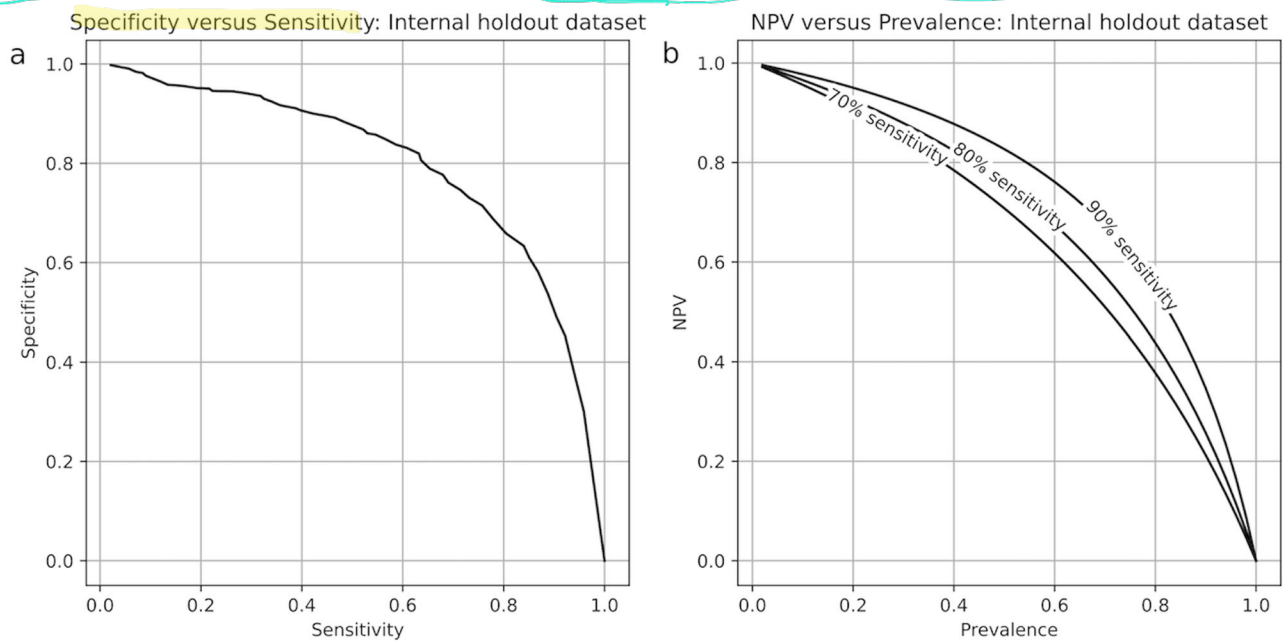
Table 2 | Model performance for detecting an elevated mPCWP ± standard error of the mean

Data Subsets	mPCWP > 18 mmHg task AUROC (*N* samples)	
	Internal-holdout dataset	**External-validation dataset**
Full dataset	0.80 ± 0.0003 (1349)	0.76 ± 0.0001 (4620)
Trustworthy	0.80 ± 0.0004 (1221)	0.76 ± 0.0001 (4273)
Untrustworthy	0.52 ± 0.006 (128)	0.52 ± 0.002 (347)

The AUROC is calculated for the full dataset, as well as for the "trustworthy" and "untrustworthy" subsets of the predictions where the "trustworthiness" metric is defined as the Shannon entropy of the corresponding prediction. Standard errors are computed over 1000 stratified bootstraps. *mPCWP* mean pulmonary capillary wedge pressure, *AUROC* area under receiver operating characteristic.

Fig. 2 | Model performance on internal-holdout set. a Calculated specificity as a function of the sensitivity using the internal-holdout set. **b** Negative predictive value (NPV) as a function of the baseline prevalence of elevated wedge in the underlying population (i.e., pre-test probability).

development set, the sensitivity is 81 percent, and the specificity is 66 percent in the internal-holdout dataset. The range of calculated specificities and sensitivities is shown in Fig. 2a.

Using these calculated sensitivities and specificities, we computed positive and negative predictive values. As predictive values are a function of the underlying prevalence of an elevated mPCWP in the population of interest, we computed predictive values as a function of different prevalence values. The PPV of the model is generally low and is only above 90 percent when the population prevalence (or pre-test probability) is high; e.g., the PPV is 75.3 percent when the pre-test probability is 50 percent (sensitivity 70 percent).

The NPV as a function of pre-test probability is shown in Fig. 2b. When the sensitivity is 80 percent and the pre-test probability is 10 percent, the NPV is 96.6 percent. Yet higher NPVs are attained at higher sensitivities. Exact results for sensitivity thresholds of 70 and 80 percent and for pre-valence values of 10 percent, 50 percent, and the observed prevalence are given in the Supplementary Results in Table S4.

Evaluation on external-validation set
We further evaluated model performance on an external-validation dataset. For this cohort, the AUROC for detecting a mPCWP > 18 mmHg was 0.76 (Table 2). We also calculated the same trustworthiness metric for the model predictions on this dataset for the mPCWP task, as described in the previous subsection. The AUROC computed specifically for the more trustworthy predictions was 0.76, and the AUROC computed for the less trustworthy predictions was 0.52 for the task of identifying an elevated mPCWP.

We again computed the sensitivities and specificities for the model for the external-validation dataset, using the same decision thresholds as for the internal-holdout dataset (Fig. 3a). The 70 percent sensitivity threshold (from the development dataset) yields a sensitivity of 55 percent in the external-validation dataset and a specificity of 82.3 percent, while the 80 percent sensitivity threshold (from the development dataset) results in a sensitivity of 68 percent and a specificity of 75 percent. The confusion matrix for the 70 percent sensitivity decision threshold is available in the Supplementary Results (Table S3).

Model performance on the external-validation set in terms of PPV and NPV are given in Supplementary Table S4, for prevalence values of 10 percent and 50 percent and for the observed prevalence in the dataset. The

PPV is over 70 percent for a prevalence of 50 percent at sensitivity thresholds of both 70 and 80 percent. The NPV exceeds 90 percent at a prevalence of 10 percent for both sensitivity thresholds reported here. Yet higher NPVs are attained at higher sensitivity thresholds, such as 90 percent. NPV as a function of pre-test probability is shown in Fig. 3b.

Model performance within diagnostic subtypes
We evaluated model performance within diagnostic subgroups. In the HF & Transplant cohort, an important subtype in the context of advances HF care, AUROC is slightly higher as compared to the entire dataset. In the internal-holdout dataset, the AUROC is 0.82 in this subgroup, and the AUROC was 0.78 in the external-validation dataset (Table 3).

CHAIS performance on ECG patch-monitor data
CHAIS was prospectively evaluated on the ECG patch-monitor dataset obtained from patients who wore a commercially available ECG patch-monitor prior to cardiac catheterization. Here, as before, our goal was to determine if the model can discriminate between patients who had an elevated mPCWP from those who do not.

In this study, the time between the last recorded ECG measurement and the actual time of the catheterization varies from patient to patient because several RHCs occurred later than their scheduled time. Delays in the timing of the RHC happen when other more urgent, unscheduled, catheterizations need to be performed first (e.g., STEMI), or due to a lack of sufficient staffing to perform non-urgent cases. We therefore evaluated model performance relative to the time between the recorded ECG and the time of the actual catheterization. Generally, discriminatory performance increases as the time to catheterization decreases (Fig. 4). Evaluating time points where 10 more samples were available, the highest AUROCs are obtained when the ECG data are acquired within 2 h of the catheterization procedure, with the best AUROC (0.875) at 1 h and 25 min before catheterization. The number of samples available at each time points are shown in Supplementary Fig. S4.

Intra-patient model performance
We explored the dynamic nature of CHAIS outputs by examining trends in model predictions within the data for individual patients. For the internal-holdout dataset, several examples are shown in Supplementary Fig. S2. For
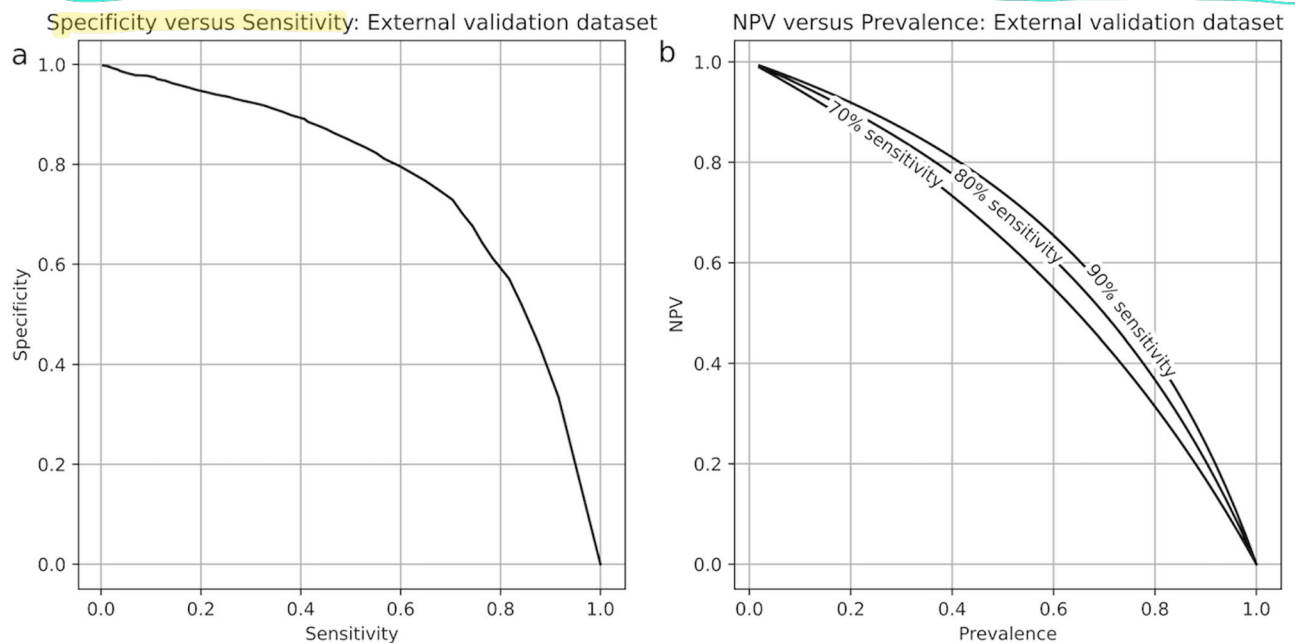
Fig. 3 | Model performance on external-validation holdout set. a Calculated specificity as a function of the sensitivity using the external-validation set. **b** Negative predictive value (NPV) as a function of the baseline prevalence of elevated wedge in the underlying population (i.e., pre-test probability).

patients where serial samples were available, predictions appear to track mPCWP on the scale of weeks. In the wearable ECG patch-monitor dataset, we observe significant changes in model outputs on the scale of hours (Supplementary Fig. S3). These findings cannot be explained solely by changes in heart rate, or using a simple model of HF, as demonstrated in our prior work[15].

Discussion

CHAIS leverages single-lead ECG data to identify patients who have an elevated mPCWP. As the mPCWP rises before the onset of symptoms, non-invasive methods that identify when the mPCWP is elevated enable early identification of patients at risk of developing symptoms of congestive HF.

CHAIS was developed and validated using Lead-I ECGs derived from in-hospital 12-lead recordings and tested on prospectively acquired wearable ECG data using a commercially available ECG patch-monitor. CHAIS exhibited good discriminatory performance for detecting elevated mPCWP in the internal-holdout (0.80) and external-validation datasets (0.76). Calculated predictive values suggest that NPVs arising from the model are informative. The NPV is greater than 95 percent when the pre-test probability is below 10 percent (at a sensitivity of 80 percent or higher) in the external-validation set, suggesting that the model can help rule out an elevated mPCWP in low-risk patients.

To determine how the model would perform on data from an ECG patch-monitor device, we prospectively studied patients who wore a commercially available patch-monitor and who underwent RHC ~24 h after monitor placement. CHAIS results on this prospective dataset were obtained in a true zero-shot fashion, with no fine-tuning on the ECG patch-monitor data. Our ultimate goal was to determine if an ECG derived from patch-monitor data can be used to estimate the coincident mPCWP. Consequently, we evaluated whether CHAIS can identify elevated mPCWP when the time between the ECG signal recording and the RHC is minimal, and explore how the discriminatory ability changes as a function of time between the ECG data collection and RHC. We found that the discriminatory ability of CHAIS increased as the time between the ECG recording and the RHC decreased. The AUROC was 0.875 when the time difference between the ECG and the RHC was 1 h and 25 min (the first time where at least 10 data points were available). We were unable to derive statistically robust estimates of the AUROC at shorter time intervals due to the paucity of patients that had ECG at times even closer to RHC. Given the relatively small number of patients in our prospective study, we were also unable to compute reliable sensitivities, specificities, and predictive values from these wearable ECG data.

To put the CHAIS AUROC scores into perspective, we note that several studies have estimated the discriminatory ability of echocardiographic measurements for identifying when the mPCWP > 18 mmHg. In particular, the ratio of the mitral inflow velocity (E wave) and tissue Doppler mitral annular velocity (e' wave)—quantities frequently measured in echocardiographic studies of patients with HF—has been proposed as a robust method for estimating when the mPCWP is elevated[32]. The AUROC of the E/e' ratio for identifying when the mPCWP > 18 mmHg has been calculated using small datasets, each containing fewer than 92 patients, yielding AUROCs between 0.68 and 0.78[33–36]. Our method has a discriminatory ability that is on par or better than these estimates and does not require a skilled practitioner to obtain Doppler velocities from cardiac ultrasound.

ECG data from our retrospective cohorts correspond to high-quality Lead I ECG signals. However, the signal quality from wrist-worn ECG devices can be poor relative to the Lead I ECG[23]. By contrast, bipolar single-

Table 3 | AUROC within subpopulation in the internal-holdout dataset for detecting elevated mPCWP

Patient Subgroups	AUROC (number of samples)	
	Internal-holdout set	External-validation set
HF & transplant	0.82 ± 0.0005 (979)	0.78 ± 0.0001 (3662)
Heart failure	0.66 ± 0.001 (552)	0.62 ± 0.0004 (1340)
Post cardiac transplant	0.57 ± 0.002 (427)	0.69 ± 0.0004 (2322)
Coronary artery disease	0.71 ± 0.003 (163)	0.65 ± 0.002 (390)
Pulmonary hypertension	0.71 ± 0.009 (62)	0.69 ± 0.01 (57)
Valvular disease	0.63 ± 0.01 (60)	0.68 ± 0.005 (122)
Other pulmonary disease	0.83 ± 0.01 (37)	0.65 ± 0.004 (226)
Other	0.94 ± 0.007 (31)	0.80 ± 0.004 (114)

Subgroups with more than 10 samples in both datasets, and at least 1 point in each class, are included (HF = heart failure). Error values correspond to the standard error of the mean.
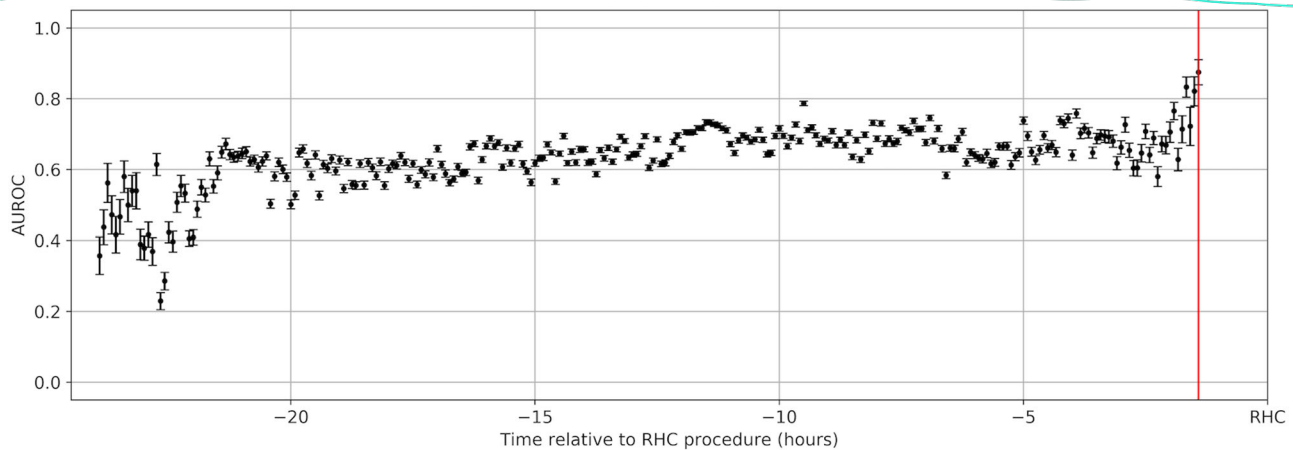
Fig. 4 | CHAIS performance on prospective data. Model performance (AUC) on the elevated mPCWP detection task versus time relative to catheterization for the ECG patch-monitor data. Samples are extracted from patients where data of sufficient quality is available at a certain time before the catheterization procedure. The plot shows AUROC versus time relative to the catheterization procedure. Error bars correspond to the standard error of the mean and are computed over 1000 stratified bootstraps. The vertical line marks 1 h and 25 min before catheterization, the time where the best AUROC is observed. AUCs are reported for time points that have 10 or more observations.

lead wearable ECG patch-monitors, which are applied to the chest wall, typically yield signals comparable to what is obtained with multi-lead high-quality ECG devices[37,38]. Consequently, our prospective study focused on a patch-monitor applied to the chest, rather than a smartwatch-acquired ECG signal. How these data generalize to signals acquired with wrist-worn monitors remains to be explored. Another limitation of our study is that the prospective evaluation relied on a small cohort of patients admitted for an RHC. Whether these results generalize to ambulatory patients, is therefore not straightforward and requires further prospective validation in larger cohorts. Nevertheless, these results show the feasibility of using patch-monitor data for invasive hemodynamic monitoring. Lastly, as our studies were not designed to evaluate whether the model is performant with respect to tracking time-dependent mPCWP changes in individual patients, the ability of the method to track mPCWP over long periods of time is unclear. Nonetheless, our results argue that this approach provides a useful screening tool that may help physicians non-invasively assess the mPCWP in symptomatic patients.

Deep learning models are not inherently interpretable, in part because of the large number of parameters associated with the models and the nonlinearity of the relationships they are trained to capture. Nevertheless, interpretability is important because it helps users gauge when to trust model predictions, and, in the context of clinical medicine, predictions that agree with one's prior understanding of pathophysiology are more likely to be trusted. We, therefore, propose a trust score, which reflects model confidence, and find that more trustworthy predictions are associated with better discriminatory ability for all the tasks we investigate and in all the datasets we examine.

This study describes a deep learning-based method to detect abnormal cardiac hemodynamics from non-invasive ECG patch-monitors. The ability to monitor cardiac hemodynamics in a non-invasive manner would be a transformative addition to the HF management toolbox.

Data availability
Access to de-identified clinical data is only available pending IRB approval. Numerical values underlying figures are included in the supplementary data files and labeled as "Source Data".

Code availability
Model architecture and weights are available online[39].

References
1. Savarese, G. et al. Global burden of heart failure: a comprehensive and updated review of epidemiology. *Cardiovasc. Res.* **118**, 3272–3287 (2022).
2. Dunlay, S. M. et al. Lifetime costs of medical care after heart failure diagnosis. *Circ. Cardiovasc. Qual. Outcomes* **4**, 68–75 (2011).
3. Khan, M. S. et al. Trends in 30- and 90-day readmission rates for heart failure. *Circ. Heart Fail.* **14**, e008335 (2021).
4. Michalsen, A., König, G. & Thimme, W. Preventable causative factors leading to hospital admission with decompensated heart failure. *Heart* **80**, 437–441 (1998).
5. Connolly, D. C., Kirklin, J. W. & Wood, E. H. The relationship between pulmonary artery wedge pressure and left atrial pressure in man. *Circ. Res.* **2**, 434–440 (1954).
6. Flores, E. D., Lange, R. A. & Hillis, L. D. Relation of mean pulmonary arterial wedge pressure and left ventricular end-diastolic pressure. *Am. J. Cardiol.* **66**, 1532–1533 (1990).
7. Aalders, M. & Kok, W. Comparison of hemodynamic factors predicting prognosis in heart failure: a systematic review. *J. Clin. Med.* **8**, 1757 (2019).
8. Mascherbauer, J. et al. Wedge pressure rather than left ventricular end-diastolic pressure predicts outcome in heart failure with preserved ejection fraction. *JACC Heart Fail.* **5**, 795–801 (2017).
9. Abraham, W. T. & Perl, L. Implantable hemodynamic monitoring for heart failure patients. *J. Am. Coll. Cardiol.* **70**, 389–398 (2017).
10. Abraham, W. T. et al. Wireless pulmonary artery haemodynamic monitoring in chronic heart failure: a randomised controlled trial. *Lancet* **377**, 658–666 (2011).
11. Zile, M. R. et al. Transition from chronic compensated to acute decompensated heart failure: pathophysiological insights obtained from continuous monitoring of intracardiac pressures. *Circulation* **118**, 1433–1441 (2008).
12. Nagueh, S. F., Middleton, K. J., Kopelen, H. A., Zoghbi, W. A. & Quiñones, M. A. Doppler tissue imaging: a noninvasive technique for evaluation of left ventricular relaxation and estimation of filling pressures. *J. Am. Coll. Cardiol.* **30**, 1527–1533 (1997).
13. Tolia, S., Khan, Z., Gholkar, G. & Zughaib, M. Validating left ventricular filling pressure measurements in patients with congestive heart failure: CardioMEMS™ pulmonary arterial diastolic pressure versus left atrial pressure measurement by transthoracic echocardiography. *Cardiol. Res. Pract.* **2018**, 8568356 (2018).

14. Drazner, M. H. et al. Value of clinician assessment of hemodynamics in advanced heart failure. *Circ. Heart Fail.* **1**, 170–177 (2008).

15. Schlesinger, D. E. et al. A deep learning model for inferring elevated pulmonary capillary wedge pressures from the 12-lead electrocardiogram. *JACC Adv.* **1**, 1–11 (2022).

16. Lindenfeld, J. et al. Haemodynamic-guided management of heart failure (GUIDE-HF): a randomised controlled trial. *Lancet* **398**, 991–1001 (2021).

17. Bumgarner, J. M. et al. Smartwatch algorithm for automated detection of atrial fibrillation. *J. Am. Coll. Cardiol.* **71**, 2381–2388 (2018).

18. Perez, M. V. et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N. Engl. J. Med.* **381**, 1909–1917 (2019).

19. Tison, G. H. et al. Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA Cardiol.* **3**, 409–416 (2018).

20. Caillol, T. et al. Accuracy of a smartwatch-derived ECG for diagnosing bradyarrhythmias, tachyarrhythmias, and cardiac ischemia. *Circ. Arrhythmia Electrophysiol.* **14**, e009260 (2021).

21. Castelletti, S. et al. A wearable remote monitoring system for the identification of subjects with a prolonged QT interval or at risk for drug-induced long QT syndrome. *Int. J. Cardiol.* **266**, 89–94 (2018).

22. Attia, Z. I. et al. Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction. *Nat. Med.* **28**, 2497–2503 (2022).

23. Funston, R. et al. Comparative study of a single lead ECG in a wearable device. *J. Electrocardiol.* **74**, 88–93 (2022).

24. Quartieri, F. et al. Artificial intelligence augments detection accuracy of cardiac insertable cardiac monitors: results from a pilot prospective observational study. *Cardiovasc. Digit. Health J.* **3**, 201–211 (2022).

25. Raghu, A. et al. ECG-guided non-invasive estimation of pulmonary congestion in patients with heart failure. *Sci. Rep.* **13**, 3923 (2023).

26. Guzzetti, S. et al. Different spectral components of 24 h heart rate variability are related to different modes of death in chronic heart failure. *Eur. Heart J.* **26**, 357–362 (2004).

27. Nohria, A. et al. Clinical assessment identifies hemodynamic profiles that predict outcomes in patients admitted with heart failure. *J. Am. Coll. Cardiol.* **41**, 1797–1804 (2003).

28. Tenny, S. & Hoffman, M. R. *StatPearls* (StatPearls Publishing, 2023).

29. Chinitz, J. S. et al. Use of a smartwatch for assessment of the QT interval in outpatients with coronavirus disease 2019. *J. Innov. Card. Rhythm Manag.* **11**, 4219–4222 (2020).

30. Strik, M. et al. Validating QT-interval measurement using the Apple Watch ECG to enable remote monitoring during the COVID-19 pandemic. *Circulation* **142**, 416–418 (2020).

31. Garabelli, P. et al. Comparison of QT interval readings in normal sinus rhythm between a smartphone heart monitor and a 12-lead ECG for healthy volunteers and inpatients receiving sotalol or dofetilide. *J. Cardiovasc. Electrophysiol.* **27**, 827–832 (2016).

32. Martelli, G. et al. Echocardiographic assessment of pulmonary capillary wedge pressure by E/e' ratio: a systematic review and meta-analysis. *J. Crit. Care* **76**, 154281 (2023).

33. Cameli, M. et al. Left atrial longitudinal strain by speckle tracking echocardiography correlates well with left ventricular filling pressures in patients with heart failure. *Cardiovasc. Ultrasound* **8**, 14 (2010).

34. Galderisi, M. et al. Site-dependency of the E/e' ratio in predicting invasive left ventricular filling pressure in patients with suspected or ascertained coronary artery disease. *Eur. Heart J. Cardiovasc. Imaging* **14**, 555–561 (2012).

35. Cowie, B., Kluger, R., Rex, S. & Missant, C. Noninvasive estimation of left atrial pressure with transesophageal echocardiography. *Ann. Card. Anaesth.* **18**, 312–316 (2015).

36. Hewing, B. et al. Left atrial strain predicts hemodynamic parameters in cardiovascular patients. *Echocardiography* **34**, 1170–1178 (2017).

37. Rajbhandary, P. L., Nallathambi, G., Selvaraj, N., Tran, T. & Colliou, O. ECG signal quality assessments of a small bipolar single-lead wearable patch sensor. *Cardiovasc. Eng. Technol.* **13**, 783–796 (2022).

38. Rosenberg, M. A., Samuel, M., Thosani, A. & Zimetbaum, P. J. Use of a noninvasive continuous monitoring device in the management of atrial fibrillation: a pilot study. *Pacing Clin. Electrophysiol.* **36**, 328–333 (2013).

39. Schlesinger, D., Alam, R. & Stultz, C. M. Website: https://github.com/mit-ccrg/CHAIS.

Author contributions
D.E.S. and C.M.S. developed the method. D.E.S. and R.A. implemented and evaluated the described methods. E.P., S.D., P.S., and J.G. provided data for the study. D.E.S., R.A., R.R., and C.M.S. contributed to the writing and editing of this article. C.M.S. supervised the work.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-024-00730-5.

Correspondence and requests for materials should be addressed to Collin M. Stultz.

Peer review information *Communications Medicine* thanks Manuel Marina-Breysse and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

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