

<https://doi.org/10.1038/s41746-026-02919-5>

Bridging the gap from clinical to home ECG: quantifying and overcoming accuracy loss in AI-enabled single-lead ECG models



Chin Lin^{1,2}, Wei-Ting Liu³, Kai-Chieh Chen^{1,2}, Dung-Jang Tsai^{1,2}, Da-Wei Chang³, Tsung-Neng Tsai³, Cheng-Chung Cheng³, Chih-Yuan Lin⁴, Yuan-Hao Chen⁵, Chien-Sung Tsai⁴, Shih-Hua Lin⁶ & Chin-Sheng Lin^{1,2,3,7} ✉

AI-enabled electrocardiogram (ECG) models trained on 12-lead recordings may underperform on home single-lead devices due to device and context domain shift. We trained Lead-I models using 676,192 ECGs from 246,874 patients and externally validated them in 219,231 ECGs from 65,338 patients. We evaluated two methodological control indicators (gender and age) and seven clinical phenotypes [90-day mortality, left ventricular ejection fraction (EF), pulmonary artery systolic pressure (PASP), left atrial diameter, N-terminal pro-B-type natriuretic peptide (NT-proBNP), hemoglobin, and estimated glomerular filtration rate]. Cross-hospital validation showed minimal overall degradation (AUC decline <0.03). In contrast, direct deployment to consumer ECGs from Apple Watch ($n = 9835$) and QOCA ECG102D ($n = 31,517$) produced substantial accuracy losses. After fine-tuning, AUCs across eight classification tasks for the Apple Watch and QOCA ECG102D were significantly improved, with high performance for low EF (0.85/0.86) and more modest performance for elevated PASP (0.73/0.68) and anemia (0.73/0.69). Performance was higher in sinus rhythm than atrial fibrillation. Learning curves indicated that approximately 500–1000 labeled ECGs are required for reliable fine-tuning; smaller sets can be harmful. We release an evaluation-only HOME dataset to facilitate reproducible benchmarking of cross-device generalization. Models should be interpreted as screening or triage tools, not standalone diagnostic tests.

The trend of sustainable chronic disease care is increasingly shifting from clinics to the home setting¹, yet most homes lack the comprehensive physiological monitoring capabilities found in hospitals or specialized centers. At present, readily accessible devices in homes primarily measure blood pressure, skin temperature, oximetry, or respiratory rate². One important class of emerging home-based monitoring technologies is single-lead electrocardiography (ECG) devices, which are typically activated by a thumb or

fingertip press. Representative examples include the Apple Watch ECG feature, AliveCor KardiaMobile, and QOCA ECG102D. These consumer devices are now widely adopted and have substantially enhanced the ability to collect ECG data outside of clinical environments. Currently, these home-based ECG devices are predominantly employed for the detection of atrial fibrillation (AF)^{3–5}. Expanding the utilization of these home ECG datasets beyond AF detection may further drive innovations in patient care.

¹Medical Technology Education Center, School of Medicine, College of Medicine, National Defense Medical University, Taipei, Taiwan, ROC. ²Military Digital Medical Center, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, ROC. ³Division of Cardiology, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, ROC. ⁴Division of Cardiovascular Surgery, Department of Surgery, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, ROC. ⁵Department of Neurological Surgery, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, ROC. ⁶Division of Nephrology, Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical University, Taipei, Taiwan, ROC. ⁷Department of Cardiology, School of Medicine, College of Medicine, National Defense Medical University, Taipei, Taiwan, ROC.

✉ e-mail: littlelincs@gmail.com

With recent advances in deep learning techniques, multiple research groups have demonstrated that artificial intelligence (AI) analysis of resting 12-lead ECGs can uncover diseases and physiological indicators undetectable by human interpreters^{6,7}, such as low ejection fraction⁸, paroxysmal AF⁹, and long-term mortality risk¹⁰. Subsequent clinical trials have shown that identifying these subtle risks in hospital settings and prompting physicians to initiate appropriate interventions may improve downstream clinical outcomes^{11–16}. Such findings raise the compelling possibility that early detection of these subtle physiological changes at home could substantially enhance current healthcare standards. Furthermore, several research teams have explored the extraction of fewer leads from standard 12-lead ECGs, demonstrating that AI can accurately predict a wide range of physiological conditions, even using solely Lead-I ECG data^{17–27}. Beyond individual feasibility studies, the 2021 PhysioNet/Computing in Cardiology Challenge (“Will two do?”) provided a large-scale community benchmark for reduced-lead ECG analysis, evaluating algorithms across 12-, 6-, 4-, 3-, and 2-lead configurations for identifying cardiac abnormalities^{28,29}. Given the growing availability of home-based ECG products and the proven capabilities of AI in ECG analysis, integrating AI with home-based ECG devices holds the potential to significantly revolutionize current healthcare practices.

However, only a limited number of research groups have truly attempted to apply AI algorithms to home-based ECG devices for detecting conditions beyond AF. For example, researchers from the Mayo Clinic enrolled over 2400 patients in the United States to perform Apple Watch ECG measurements. Within a subset of approximately 420 patients, they demonstrated that AI prediction of low ejection fraction achieved an area under the curve (AUC) of about 0.88³⁰. Apart from this notable example, only a few studies have rigorously collected data from home-based ECG devices to validate the performance of AI models using relatively small sample sizes^{31–33}. Although many feasibility studies have indicated that fewer ECG leads can still provide accurate disease prediction^{17–24,26,27}, AI models often lose accuracy when deployed across hospitals, with even greater degradation across devices. Studies have shown high rates of poor signal quality among data recorded by home-based ECG devices, and significant variation in quality across different manufacturers³⁴. Additionally, studies have reported that the patient’s posture during ECG measurement can significantly alter heart rate³⁵, and home-based ECG acquisition methods differ markedly from traditional resting 12-lead ECG procedures. These discrepancies highlight the urgent need for large-scale cohort studies to fully explore the potential of AI applications on home-based ECG devices.

In this context, the present study utilizes a uniquely large and diverse dataset of single-lead ECG recordings from home-use devices (including the Apple Watch and the QOCA ECG102D) to predict a wide range of patient phenotypes. The primary objectives of this study include: (1) evaluating whether AI models trained using fewer leads extracted from resting 12-lead ECGs can be directly applied to home-based ECG devices, and quantifying the extent of accuracy loss; and (2) if fine-tuning is required, determining the extent to which accuracy can be restored and establishing how many training samples are necessary for effective model refinement. The findings from this investigation will provide critical guidance for future developments in home-based healthcare.

Results

Patient characteristics across ECG datasets

The datasets of this study are summarized in Fig. 1. ECG patterns, patient demographics, clinical characteristics, echocardiographic findings, and laboratory results are detailed in Table 1. Across datasets, patients had similar age distributions (approximately 58–63 years), with a slightly higher male representation (52.2–53.5%). Phenotypes were linked to each ECG using pre-specified time windows based on clinical relevance: echocardiography within ± 7 days³⁶; N-terminal pro B-type natriuretic peptide (NT-proBNP) within ± 24 h³⁷; hemoglobin (Hb) within ± 12 h³⁸; and estimated glomerular filtration rate (eGFR) within ± 30 days³⁹. Because echocardiography and laboratory testing were performed as part of routine care, label

availability varied substantially by phenotype and device cohort. Therefore, each prediction task used only ECGs with an eligible label within the pre-defined window, thereby reducing the effective sample size for certain outcomes (e.g., NT-proBNP) and potentially increasing uncertainty in performance estimates. Echocardiographic parameters (left ventricular ejection fraction [EF], pulmonary artery systolic pressure [PASP], left atrial diameter [LAD]) and laboratory measures (NT-proBNP, Hb, eGFR) showed comparable distributions between hospital-based and home-based ECG datasets, though home-based ECG cohorts had slightly healthier echocardiographic profiles (higher average EF and smaller heart chamber dimensions) and better survival rates on average. The distribution of trainable clinical labels across four ECG cohorts (internal 12-lead ECG, external 12-lead ECG, Apple Watch, and QOCA ECG102D) is shown in Fig. S1. Kernel density plots illustrate comparable distributions for age, echocardiographic parameters, and laboratory outcomes, highlighting generally overlapping distributions among cohorts, with minor shifts towards healthier profiles in home-based ECG groups. To facilitate model training and testing, we randomly partitioned the 12-lead ECG dataset into training, internal test, and external test subsets. Detailed distributions for these subsets are summarized in Table S1. Similarly, we partitioned home-based ECG data (Apple Watch and QOCA ECG102D) into respective training and test sets, detailed in Table S2.

ECG waveform characteristics and parameter agreement

Representative ECG waveforms and acquisition methods from the three device sources (Apple Watch, QOCA ECG102D, and Philips TC-series) are illustrated in Fig. 2. Visually, the waveforms from the home-based ECG devices display more prominent baseline wander and noise compared with clinical-grade ECG recordings, despite additional filtering. In a subset of 157 individuals, each of whom underwent all three recordings (Apple Watch, QOCA ECG102D, and Philips 12-lead) within a 10-min window, Bland-Altman plots show agreement between heart rates derived from the Apple Watch and QOCA ECG102D vendor-provided algorithms compared with the Philips TC-series ECG reference. The heart-rate differences were 7 (95% range: -9 to 22) for Apple Watch and 12 (95% range: -14 to 37) for QOCA ECG102D, respectively. Using the same paired subset, we further evaluated agreement in commonly used ECG timing parameters beyond heart rate, including PR interval, QRS duration, and QTc interval. As shown in Fig. S2, these interval measurements exhibited device-specific biases and variability, with larger systematic deviations observed for QOCA ECG102D than for Apple Watch. Collectively, these findings indicate that cross-device differences extend beyond heart rate and can also affect fundamental waveform timing characteristics relevant to downstream AI inference.

Methodological experiments of base and fine-tuned Lead-I ECG models

In Fig. 3, we first evaluated the generalizability of the base Lead-I ECG AI models in classification tasks, trained on internal clinical ECG data. Cross-site validation using an external hospital dataset demonstrated only minor performance degradation across most prediction tasks, indicating robust model generalization within hospital-acquired ECG datasets. However, substantial accuracy loss was observed when directly applying these base models to home-based ECG recordings from Apple Watch and QOCA ECG102D, highlighting significant cross-device domain shifts. Notably, the accuracy drop was more pronounced in the QOCA ECG102D dataset compared to the Apple Watch. Taking the detection of low EF ($\leq 40\%$) as an example, the base model performance decreased considerably from an internal test AUC of approximately 0.88 to around 0.82 in the Apple Watch dataset, and further declined to 0.75 in the QOCA ECG102D dataset. Similar trends of greater accuracy loss on the QOCA ECG102D device were observed across most other phenotypes. Importantly, fine-tuning models using a small number of device-specific ECG samples dramatically restored predictive accuracy across nearly all phenotypes, especially for the QOCA ECG102D device. After fine-tuning, the low EF detection task achieved

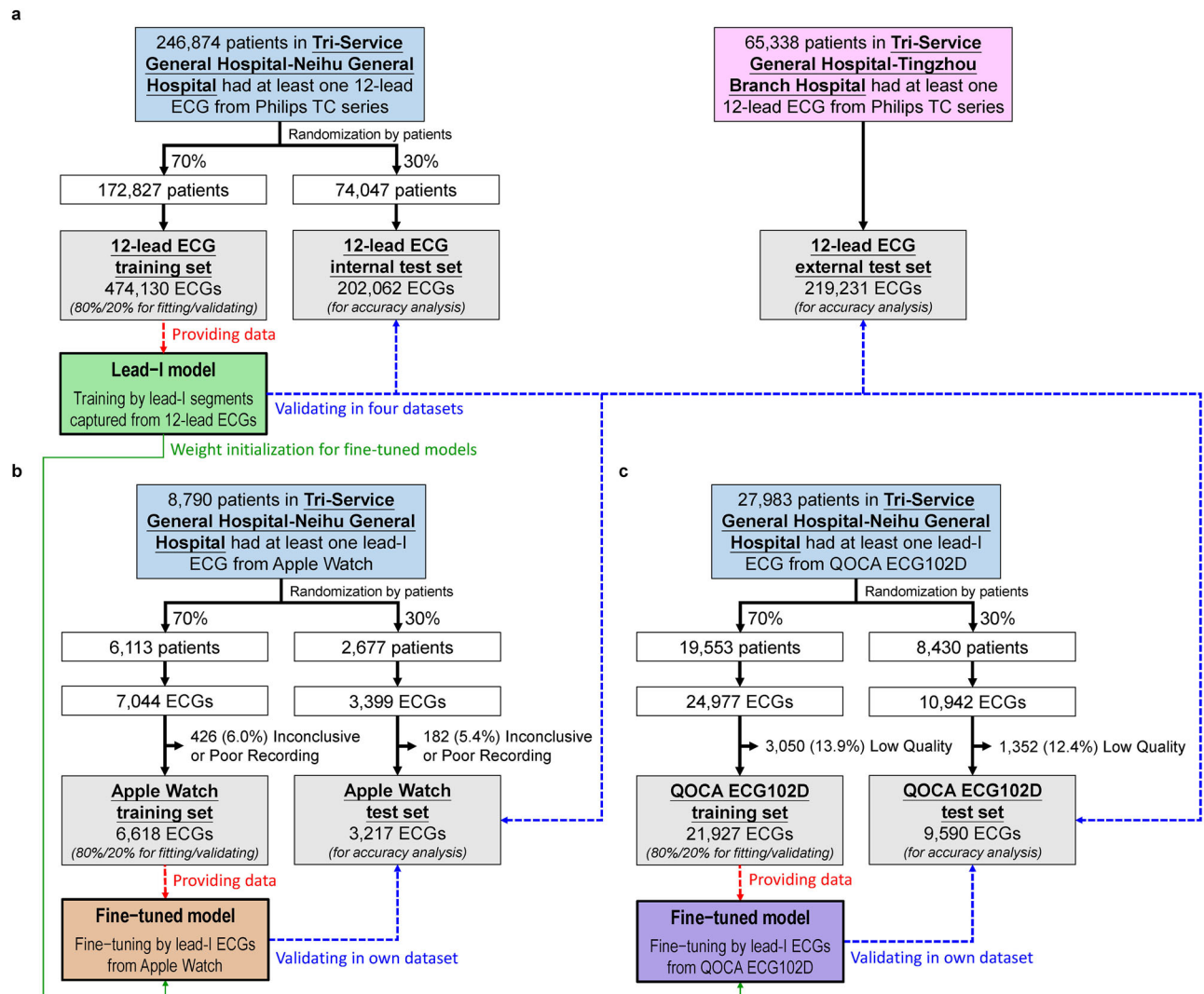


Fig. 1 | Study design and dataset workflow for training, fine-tuning, and testing of Lead-I ECG AI models. **a** Development of the base Lead-I model: 246,874 patients at Tri-Service General Hospital-Neihu General Hospital were randomized at a 70%/30% ratio into a training set (172,827 patients, 474,130 ECGs) and an internal test set (74,047 patients, 202,062 ECGs). An independent external test set comprised 65,338 patients (219,231 ECGs) from the Tingzhou Branch Hospital. Lead-I segments extracted from the 12-lead ECGs were used to train and validate the base model. **b** Fine-tuning on Apple Watch data: 8790 patients with single-lead Apple Watch recordings at the Neihu site were randomized 70%/30% into a training set (6113 patients, 7044 ECGs) and a test set (2677 patients, 3399 ECGs). ECG recordings judged as inconclusive or poor quality were excluded, resulting in 6618 ECGs for training and 3217 ECGs for testing. **c** Fine-tuning on QOCA ECG102D data: 27,983 patients with single-lead QOCA ECG102D recordings at the Neihu site were randomized 70%/30% into a training set (19,553 patients, 24,977 ECGs) and a test set

(8430 patients, 10,942 ECGs). After excluding low-quality ECG recordings, 21,927 ECGs remained in the training set and 9590 ECGs in the test set. Among the three training sets, we allocated 80% for model fitting and 20% to select the best-performing model based on minimal loss. Moreover, the threshold for receiver operating characteristic curve analysis was also selected from this 20% validating set. All cohort-level fitting/validating/test partitions were performed at the patient level. Lead-I segment cohort consisted of 12-lead ECGs acquired before February 2022, whereas the Apple Watch and QOCA ECG102D cohorts were composed predominantly of recordings obtained after January 2023. Because the data were collected through longitudinal routine clinical care, some patients appeared in more than one cohort. However, patients appearing in multiple cohorts were always assigned to the same partition across cohorts; therefore, no patient in a development partition could appear in a test partition of another cohort.

marked improvement, with AUC values increasing from 0.75 to around 0.86 in the QOCA ECG102D dataset, surpassing the improvement observed in the Apple Watch dataset. Receiver operating characteristic (ROC) curves for all eight classification tasks on the Apple Watch and QOCA ECG102D test sets are presented in Fig. S3.

For continuous prediction tasks (Fig. 4a), Pearson's correlation coefficients showed consistent improvements with fine-tuning. For instance, the prediction of patient age and EF demonstrated notable accuracy recovery after fine-tuning. Specifically, correlations for EF prediction increased after fine-tuning, from 0.37 to 0.45 for Apple Watch and from 0.26 to 0.41 for QOCA ECG102D. Similar improvements were consistently observed across most tasks. In addition to Pearson's correlation coefficient, we also reported

mean absolute error (MAE) to quantify regression performance in the original clinical units and to better reflect practical prediction error magnitude (Fig. 4b). Cross-device transfer from clinical 12-lead derived Lead-I ECGs to home-based single-lead recordings led to both reduced correlations and increased MAE across multiple continuous endpoints. Importantly, device-specific fine-tuning improved both association and absolute error, resulting in consistent MAE reductions across the seven continuous targets. To further characterize agreement beyond summary statistics, we provide task-level scatter plots and Bland-Altman analyzes (bias and 95% limits of agreement) in Figs. S4 and S5. Although inverse regression correction can effectively mitigate the regression-to-the-mean (Fig. S6), but at the cost of a significant increase in MAE. Single-lead AI-ECG appears better suited for

Table 1 | Distribution of labels available for each database

	12-lead ECG at Neihu (<i>n</i> = 676,192)	12-lead ECG at Tingzhou (<i>n</i> = 219,231)	Apple Watch (<i>n</i> = 9835)	QOCA ECG102D (<i>n</i> = 31,517)
ECG pattern [‡]				
Sinus rhythm	603,105 (89.2%)	191,680 (87.4%)	9188 (93.4%)	29,476 (93.5%)
Atrial fibrillation	37,791 (5.6%)	15,030 (6.9%)	647 (6.6%)	2041 (6.5%)
Others	35,296 (5.2%)	12,521 (5.7%)		
Gender				
Female	314,382 (46.5%)	104,724 (47.8%)	4883 (49.6%)	14,806 (47.0%)
Male	361,810 (53.5%)	114,507 (52.2%)	4952 (50.4%)	16,711 (53.0%)
Age (mean ± standard deviation)	59.7 ± 17.7	62.8 ± 18.2	61.3 ± 16.3	58.1 ± 16.7
Death in 90 days	<i>n</i> = 543,649 (80.4%)	<i>n</i> = 192,067 (87.6%)	<i>n</i> = 8316 (84.6%)	<i>n</i> = 26,552 (84.2%)
Survival	517,200 (95.1%)	185,111 (96.4%)	8216 (98.8%)	26,316 (99.1%)
Death	26,449 (4.9%)	6956 (3.6%)	100 (1.2%)	236 (0.9%)
Ejection fraction	<i>n</i> = 142,430 (21.1%)	<i>n</i> = 47,308 (21.6%)	<i>n</i> = 7039 (71.6%)	<i>n</i> = 8959 (28.4%)
mean ± standard deviation (%)	62.1 ± 13.4	62.9 ± 12.7	64.4 ± 9.3	64.3 ± 9.2
>40%	129,390 (90.8%)	43,640 (92.2%)	6758 (96.0%)	8631 (96.3%)
≤40%	13,040 (9.2%)	3668 (7.8%)	281 (4.0%)	328 (3.7%)
Pulmonary artery systolic pressure	<i>n</i> = 141,910 (21.0%)	<i>n</i> = 47,113 (21.5%)	<i>n</i> = 7034 (71.5%)	<i>n</i> = 8957 (28.4%)
mean ± standard deviation (mmHg)	34.0 ± 11.9	34.2 ± 12.1	32.1 ± 10.9	31.3 ± 10.5
<50 mmHg	128,870 (90.8%)	42,409 (90.0%)	6585 (93.6%)	8465 (94.5%)
≥50 mmHg	13,040 (9.2%)	4704 (10.0%)	449 (6.4%)	492 (5.5%)
Left atrial diameter	<i>n</i> = 141,836 (21.0%)	<i>n</i> = 47,143 (21.5%)	<i>n</i> = 7044 (71.6%)	<i>n</i> = 8962 (28.4%)
mean ± standard deviation (mm)	39.0 ± 7.7	39.6 ± 7.8	37.4 ± 7.1	37.7 ± 7.1
<50 mm	128,674 (90.7%)	42,460 (90.1%)	6661 (94.6%)	8450 (94.3%)
≥50 mm	13,162 (9.3%)	4683 (9.9%)	383 (5.4%)	512 (5.7%)
N-terminal pro B-type natriuretic peptide	<i>n</i> = 45,201 (6.7%)	<i>n</i> = 11,700 (5.3%)	<i>n</i> = 405 (4.1%)	<i>n</i> = 1394 (4.4%)
mean ± standard deviation (pg/mL)	5663.5 ± 10285.6	6008.4 ± 10306.2	4197.7 ± 8416.6	2445.4 ± 6628.3
<1000 pg/mL	24,107 (53.3%)	5769 (49.3%)	224 (55.3%)	1048 (75.2%)
≥1000 pg/mL	21,094 (46.7%)	5931 (50.7%)	181 (44.7%)	346 (24.8%)
Hemoglobin	<i>n</i> = 463,720 (68.6%)	<i>n</i> = 135,770 (61.9%)	<i>n</i> = 1485 (15.1%)	<i>n</i> = 11,892 (37.7%)
mean ± standard deviation (g/dL)	12.8 ± 2.4	12.6 ± 2.4	12.0 ± 2.3	13.0 ± 2.1
>10 g/dL	399,530 (86.2%)	115,506 (85.1%)	1182 (79.6%)	10,811 (90.9%)
≤10 g/dL	64,190 (13.8%)	20,264 (14.9%)	303 (20.4%)	1081 (9.1%)
Estimated glomerular filtration rate	<i>n</i> = 373,020 (55.2%)	<i>n</i> = 100,864 (46.0%)	<i>n</i> = 3210 (32.6%)	<i>n</i> = 16,606 (52.7%)
mean ± standard deviation (mL/min)	76.6 ± 32.8	72.3 ± 32.7	70.5 ± 35.5	82.7 ± 29.8
>60 mL/min	271,140 (72.7%)	68,907 (68.3%)	2188 (68.2%)	13,461 (81.1%)
≤60 mL/min	101,880 (27.3%)	31,957 (31.7%)	1022 (31.8%)	3145 (18.9%)

‡: The Apple Watch and QOCA ECG102D datasets only included ECGs of sinus rhythm or atrial fibrillation identified by vendor algorithms.

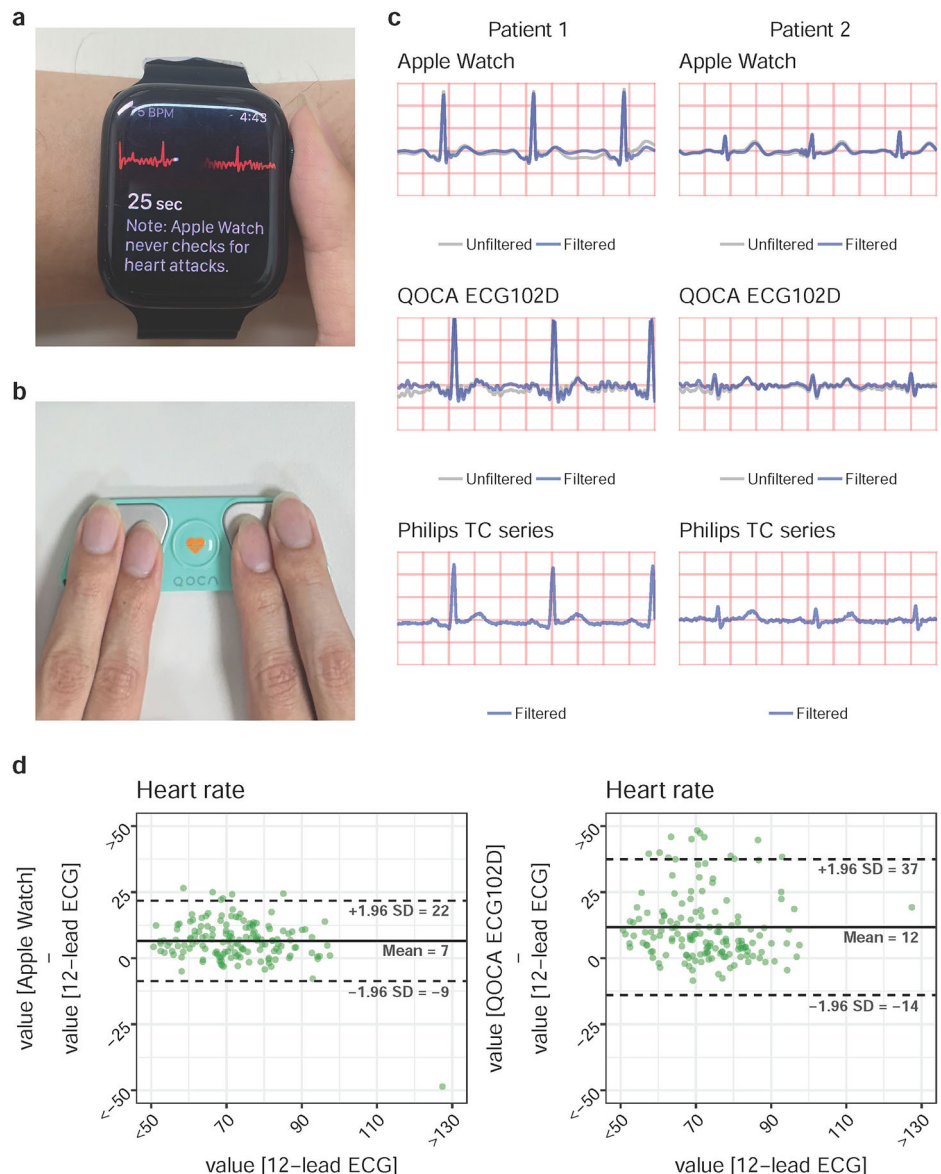
screening rather than reporting physiologic parameters, given these characteristics.

Figures S7 and S8 show the effect of additional filtering and support the validity of our additional filtering preprocessing strategy. To evaluate whether adapting only the final decision layer is sufficient, we compared full-parameter fine-tuning, which constitutes the primary approach of this study, with linear probing (frozen backbone) (Table S3). Full-parameter fine-tuning yielded equal or better performance across most endpoints, with more pronounced gains in the QOCA ECG102D cohort, consistent with a larger device-related domain shift. For example, in the QOCA ECG102D cohort, NT-proBNP classification improved from AUC 0.71 (linear probing) to 0.79 (full fine-tuning), and eGFR prediction improved from $r = 0.26$ to 0.37. In the Apple Watch cohort, linear probing was occasionally comparable for select tasks, but full fine-tuning remained the most consistent approach overall.

Screening value for portable ECG devices

To translate model discrimination into clinically interpretable risk shifts, we summarized threshold-dependent prior-to-posterior risk estimates for clinically actionable binary endpoints in Table 2. For each endpoint and device cohort, the prior risk was defined as the observed outcome prevalence in the wearable test cohort. We reported two validation-derived operating points: a sensitivity-prioritized threshold selected to target 90% sensitivity in the validation subset and a Youden-index threshold selected to provide a balanced operating point. In general, the 90%-sensitivity threshold lowered the residual risk after a negative screen but did so at the cost of lower specificity and lower positive-screen yield. For example, for EF ≤ 40%, the prior risk was 3.5% in the Apple Watch cohort and 3.9% in the QOCA ECG102D cohort. At the 90%-sensitivity threshold, positive AI-ECG screens increased the post-test risk to 12.6% and 12.0%, respectively, whereas negative screens reduced the residual risk to 0.8% and 1.0%. At the

Fig. 2 | ECG acquisition interfaces, waveform presentation, and heart-rate agreement. **a** Single-lead ECG recording on an Apple Watch worn on the wrist. **b** QOCA ECG102D single-lead device held between two fingers to record Lead-I ECG. **c** For two representative patients (Patient 1, left column; Patient 2, right column), example 10-s Lead-I traces from three devices: Apple Watch (top row), QOCA ECG102D (middle row), and Philips TC-series 12-lead system (bottom row). Gray lines are raw, unfiltered signals; blue lines represent signals after band-pass filtering and baseline correction. **d** Bland–Altman plots comparing heart rates (beats per minute) reported by each device's vendor algorithm against the Philips 12-lead ECG reference. Left: Apple Watch vs. 12-lead; right: QOCA ECG102D vs. 12-lead. Solid horizontal lines show the mean difference (bias), dashed lines show the limits of agreement (± 1.96 SD), and each green dot represents one patient's paired heart-rate difference plotted against the reference ECG value. We used the Philips 12-lead ECG as the clinical reference and therefore plotted differences against the reference value (rather than the mean of two methods) to directly visualize wearable-to-clinic domain shift; bias and limits of agreement are unaffected by this choice.



Youden-index threshold, the corresponding positive-screen post-test risks increased to 18.2% and 14.1%, while the residual risks remained at 1.2% in both cohorts. Similar trade-offs were observed across other endpoints. These results indicate that abnormal wearable AI-ECG outputs should be interpreted as prompts for confirmatory testing or clinical review, such as echocardiography for suspected low EF, elevated PASP, or enlarged LAD, and laboratory testing for suspected elevated NT-proBNP, anemia, or reduced renal function. To further examine clinical applicability across baseline-risk strata, we performed gender- and age-stratified analyses. Stratified AUCs are summarized in Table S4, and detailed threshold-dependent metrics are provided in Tables S5–S11.

Impact of underlying cardiac rhythm on AI model performance

We further explored whether underlying cardiac rhythms (sinus rhythm versus AF) affected the predictive performance of the fine-tuned AI models (Table 3). Generally, predictions derived from ECG segments recorded during sinus rhythm demonstrated superior accuracy compared to those recorded during AF. For instance, in the Apple Watch dataset, the AUC for detecting enlarged LAD was 0.76 in sinus rhythm and decreased to 0.59 in AF. A similar pattern was observed in the QOCA ECG102D dataset, where sinus rhythm ECGs provided generally higher predictive accuracy, such as an AUC of 0.73 for enlarged LAD detection, compared to 0.60 in AF.

Nevertheless, several tasks maintained robust predictive accuracy even during AF. For 90-day death prediction, we observed AUCs of 0.82 during sinus rhythm and 0.82 during AF in the Apple Watch dataset, and AUCs of 0.75 during sinus rhythm and 0.74 during AF in the QOCA ECG102D dataset. Continuous prediction tasks revealed similar findings. These results collectively emphasize the importance of rhythm-specific considerations when interpreting AI-derived ECG predictions. Although predictions generally exhibit higher accuracy in sinus rhythm, several clinically relevant tasks retain strong predictive performance even when ECGs are recorded during AF.

Given the rhythm-dependent performance differences, we further examined whether excluding AF recordings during device-specific fine-tuning could improve predictive performance for sinus rhythm ECGs. We retrained alternative fine-tuned models using sinus rhythm ECGs only in the wearable training sets and evaluated performance restricted to sinus rhythm ECGs in the corresponding wearable test sets. As summarized in Table S12, sinus rhythm-only fine-tuning did not provide consistent performance gains. These results suggest that retaining AF recordings during fine-tuning may improve generalization even for sinus rhythm evaluation, potentially by providing additional rhythm and morphology diversity and preserving informative high-risk examples that may aid model robustness.

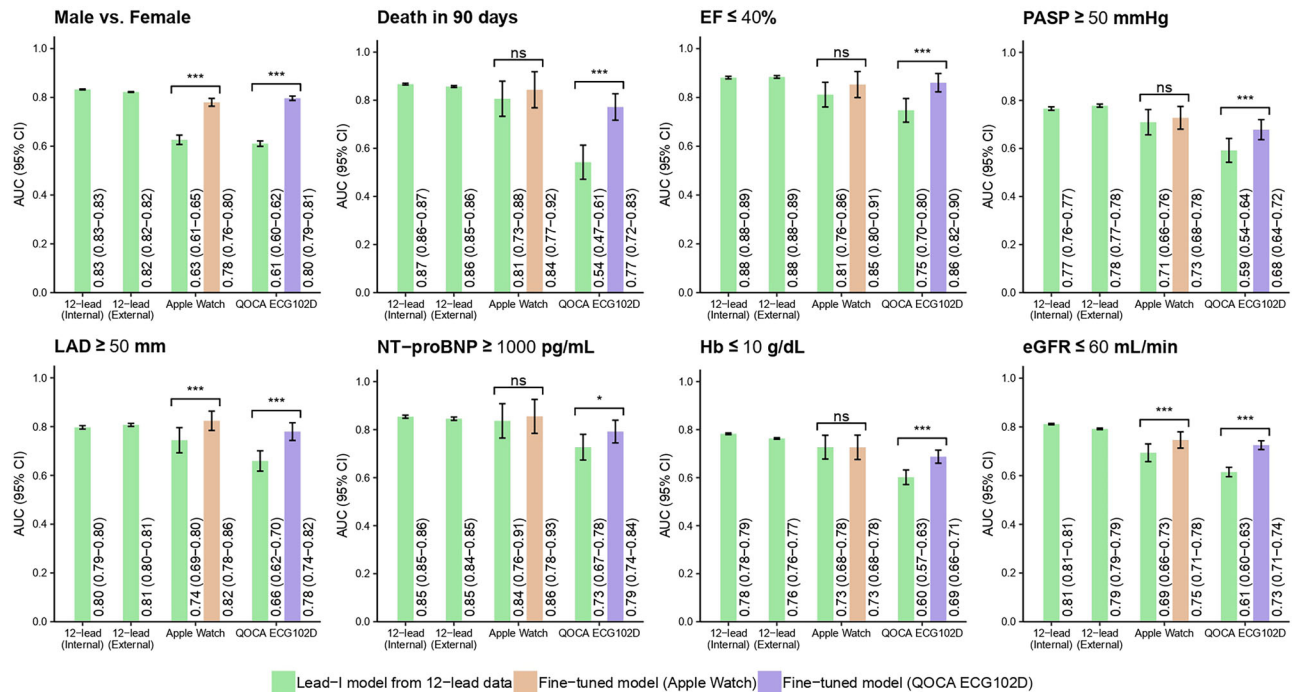


Fig. 3 | Classification performance of base and fine-tuned Lead-I ECG AI models. AUCs (95% CI) for eight clinical and demographic classification tasks using Lead-I AI-ECG models before and after device-specific fine-tuning. Green bars show the base model trained on Lead-I segments from 12-lead ECGs on each test set. Orange bars show the same model after fine-tuning on Apple Watch Lead-I ECGs (tested on

the Apple Watch cohort), and purple bars show performance after fine-tuning on QOCA ECG102D Lead-I ECGs (tested on the QOCA ECG102D cohort). Error bars denote 95% confidence intervals. Symbols indicate statistical comparisons: ns, not significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Relationship between fine-tuning sample size and prediction performance

We further analyzed the relationship between the number of ECG samples used for fine-tuning and the resulting improvements in predictive performance. Figure 5 presents the overall trend in performance change (Δ AUC for binary tasks and Δr for continuous tasks) with increasing fine-tuning dataset size. Performance rose sharply with small sample sizes but plateaued at larger ones, indicating diminishing marginal returns. It is important to note that when the sample size is very small, the fine-tuning process is actually harmful to model accuracy. Specifically, in the Apple Watch dataset, binary task performance (average Δ AUC) became comparable to the base model trained on clinical ECG data at around 630 fine-tuning samples. The QOCA ECG102D dataset required slightly smaller sample sizes (416 samples) to achieve similar performance parity. For the continuous tasks, ~388 samples were the minimal required for the fine-tuning process in both Apple Watch and QOCA ECG102D datasets. Detailed learning curves for individual tasks in the Apple Watch and QOCA ECG102D datasets are presented in Figures S9 and S10.

Released benchmark dataset

To facilitate standardized cross-device evaluation, we curated the Health records-linked Open Multi-consumer device Electrocardiogram (HOME) benchmark from the Apple Watch and QOCA ECG102D wearable test cohorts. Specifically, we sampled 1000 unique patients from the Apple Watch test set and 1000 unique patients from the QOCA ECG102D test set, retaining one ECG recording per patient. HOME therefore comprises 2000 single-lead ECG recordings, defining 18 device-endpoint benchmark tasks (the same nine clinical phenotypes evaluated separately in each device cohort). Table 4 summarizes the task list, cohort-level demographic summaries (gender and age), and task-specific sample sizes, and presents the predictive performance before (base model) and after (device-specific fine-tuning) adaptation on HOME. A detailed comparison between HOME and the full original wearable test

sets is provided in Table S13. Overall, HOME preserves the key empirical pattern observed in the full cohorts—substantial performance degradation under direct cross-device transfer and consistent recovery after device-specific fine-tuning—thereby enabling reproducible benchmarking for future methodological studies.

Discussion

In this large-scale study, we systematically quantified the accuracy loss associated with cross-device domain shifts when applying AI-enabled Lead-I ECG models—trained on resting 12-lead ECG data—to home-based ECG devices. Cross-hospital validation across two clinical sites showed minimal performance loss, whereas direct application to Apple Watch and QOCA ECG102D recordings led to substantial degradation, consistent with a marked device-related domain shift. With sufficient samples, targeted fine-tuning using device-specific ECG recordings significantly restored predictive accuracy for most clinical tasks, particularly in the QOCA ECG102D dataset. Moreover, we observed that AI models maintained clinically relevant performance even under challenging conditions such as AF, though accuracy generally remained higher during sinus rhythm.

Our findings complement the 2021 PhysioNet/Computing in Cardiology Challenge on reduced-lead ECG modeling. While that benchmark suggested that reducing the number of leads may incur only modest aggregate performance changes for some diagnostic objectives, it also demonstrated substantial performance drops when models were evaluated on hidden test sources not represented in training, underscoring the central role of dataset and domain shift in AI-ECG^{28,29}. In our setting, moving from clinical 12-lead ECG systems (from which Lead-I is extracted) to consumer-grade wearable recordings represents an even more pronounced domain shift that includes differences in hardware, electrode contact, recording posture, and real-world noise. Consistent with this, we observed minimal performance loss across hospitals within clinical ECG systems but marked degradation when directly applying the same models to consumer devices. Near-simultaneous paired recordings showed systematic differences not

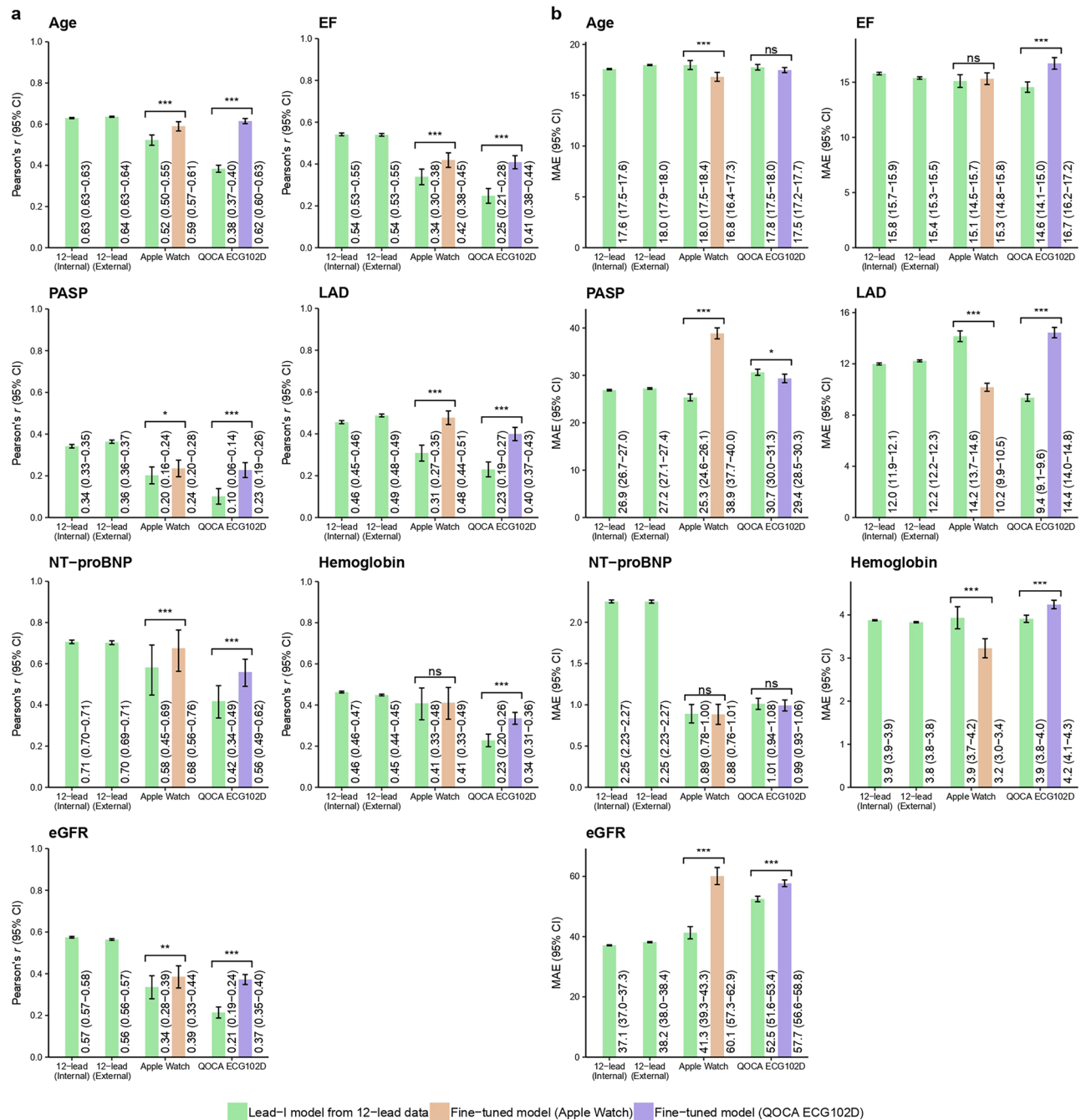


Fig. 4 | Continuous-variable prediction performance of base and fine-tuned Lead-I ECG AI models. a Continuous-variable prediction performance: Pearson's correlation coefficients (r) with 95% CIs for seven continuous clinical variables predicted by Lead-I AI-ECG models before and after device-specific fine-tuning. **b** Continuous-variable prediction performance: Mean absolute error (MAE) with 95% CIs for seven continuous clinical variables predicted by Lead-I AI-ECG models before and after device-specific fine-tuning. Green bars show the base model trained

on Lead-I segments from 12-lead ECGs on each test set. Orange bars show the same model after fine-tuning on Apple Watch Lead-I ECGs (tested on the Apple Watch cohort), and purple bars show performance after fine-tuning on QOCA ECG102D Lead-I ECGs (tested on the QOCA ECG102D cohort). Error bars denote 95% confidence intervals. Symbols indicate statistical comparisons: ns, not significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

only in heart rate but also in common timing intervals (PR and QTc), indicating that cross-device discrepancies extend beyond simple rate variation and may affect features relevant to downstream AI inference^{37,40,41}. These observations provide a plausible mechanistic explanation for significant accuracy loss. Importantly, the relevant shift is not purely technical noise. Posture-related autonomic and electrophysiologic changes can alter heart rate and repolarization timing, and prior physiologic studies have shown that postural change affects heart rate, QT-related dynamics, and the relative timing of depolarization and repolarization phases^{35,42,43}. Such

physiologic changes are difficult to reproduce with simple synthetic transformations: for example, changing heart rate by uniformly resizing the time axis would also scale QRS duration, whereas real posture-related heart-rate changes may not alter QRS duration. Although recent work has proposed synthetic or noise-adapted training as a complementary strategy for wearable ECG translation^{23,25}—these studies support synthetic noise simulation as a valuable pretraining or augmentation strategy, our results suggest that real-device fine-tuning remains necessary for deployment because it can adapt to manufacturer-specific preprocessing, electrode-contact behavior,

Table 2 | Prior-to-posterior risk interpretation of full-parameter fine-tuned AI-ECG models for clinically actionable binary endpoints

Cohort	Threshold	AUC (95% CI)	Prior risk (prevalence)	Sensitivity	Specificity	Posterior risk if positive (PPV)	Residual risk if negative (1 - NPV)
Death in 90 days							
Apple Watch	90% sensitivity	0.84 (0.77–0.92)	1.0%	89.3%	44.2%	1.6%	0.3%
Apple Watch	Youden index	0.84 (0.77–0.92)	1.0%	71.4%	79.0%	3.4%	0.4%
QOCA ECG102D	90% sensitivity	0.77 (0.72–0.83)	1.0%	88.8%	36.7%	1.4%	0.3%
QOCA ECG102D	Youden index	0.77 (0.72–0.83)	1.0%	55.0%	84.4%	3.4%	0.5%
EF ≤ 40%							
Apple Watch	90% sensitivity	0.85 (0.80–0.91)	3.5%	81.6%	79.3%	12.6%	0.8%
Apple Watch	Youden index	0.85 (0.80–0.91)	3.5%	69.7%	88.5%	18.2%	1.2%
QOCA ECG102D	90% sensitivity	0.86 (0.82–0.90)	3.9%	81.0%	75.9%	12.0%	1.0%
QOCA ECG102D	Youden index	0.86 (0.82–0.90)	3.9%	75.2%	81.4%	14.1%	1.2%
PASP ≥ 50 mmHg							
Apple Watch	90% sensitivity	0.73 (0.68–0.78)	5.7%	92.7%	24.4%	6.9%	1.8%
Apple Watch	Youden index	0.73 (0.68–0.78)	5.7%	53.7%	80.6%	14.3%	3.4%
QOCA ECG102D	90% sensitivity	0.68 (0.64–0.72)	4.9%	88.7%	34.7%	6.6%	1.7%
QOCA ECG102D	Youden index	0.68 (0.64–0.72)	4.9%	28.6%	87.5%	10.6%	4.1%
LAD ≥ 50 mm							
Apple Watch	90% sensitivity	0.82 (0.78–0.86)	5.1%	84.4%	61.1%	10.4%	1.3%
Apple Watch	Youden index	0.82 (0.78–0.86)	5.1%	78.0%	71.6%	12.8%	1.6%
QOCA ECG102D	90% sensitivity	0.78 (0.74–0.82)	6.4%	90.1%	46.0%	10.2%	1.4%
QOCA ECG102D	Youden index	0.78 (0.74–0.82)	6.4%	66.9%	73.9%	14.8%	3.0%
NT-proBNP ≥ 1000 pg/mL							
Apple Watch	90% sensitivity	0.86 (0.78–0.93)	43.6%	94.1%	60.6%	64.9%	7.0%
Apple Watch	Youden index	0.86 (0.78–0.93)	43.6%	86.3%	75.8%	73.3%	12.3%
QOCA ECG102D	90% sensitivity	0.79 (0.74–0.84)	26.2%	90.0%	41.6%	35.4%	7.9%
QOCA ECG102D	Youden index	0.79 (0.74–0.84)	26.2%	69.1%	73.2%	47.8%	13.0%
Hb ≤ 10 g/dL							
Apple Watch	90% sensitivity	0.73 (0.68–0.78)	20.9%	91.4%	42.8%	29.6%	5.0%
Apple Watch	Youden index	0.73 (0.68–0.78)	20.9%	89.2%	47.9%	31.1%	5.6%
QOCA ECG102D	90% sensitivity	0.69 (0.66–0.71)	9.7%	89.0%	36.2%	13.0%	3.2%
QOCA ECG102D	Youden index	0.69 (0.66–0.71)	9.7%	62.7%	64.3%	15.9%	5.9%
eGFR ≤ 60 mL/min							
Apple Watch	90% sensitivity	0.75 (0.71–0.78)	29.0%	89.5%	33.9%	35.6%	11.2%
Apple Watch	Youden index	0.75 (0.71–0.78)	29.0%	72.1%	66.8%	46.9%	14.5%
QOCA ECG102D	90% sensitivity	0.73 (0.71–0.74)	19.0%	88.8%	32.8%	23.6%	7.4%
QOCA ECG102D	Youden index	0.73 (0.71–0.74)	19.0%	57.2%	74.6%	34.5%	11.8%

For each endpoint and device cohort, prior risk was defined as the observed subgroup prevalence. “90% sensitivity” represents the cut point where 90% sensitivity in the validating subset; “Youden index” represents the cut point where maximum of Youden index in the validating subset. Posterior risk after a positive AI-ECG screen was defined as PPV, and residual risk after a negative AI-ECG screen was defined as 1 – NPV. The 90% sensitivity threshold was selected using the full validation subset and then applied unchanged to the test set and subgroups; therefore, observed subgroup sensitivity may differ from exactly 90%.

AUC area under receiver operating characteristic curve, PPV positive predictive value, NPV negative predictive value.

acquisition posture, and physiologic waveform changes that are unlikely to be fully captured by generic noise models.

We observed notable variability depending on both the specific device and the clinical prediction task. Overall, the accuracy degradation was less pronounced in the Apple Watch dataset compared to the QOCA ECG102D dataset. This discrepancy likely reflects the higher similarity of Apple Watch signals to Philips ECG signals, evidenced by visibly lower noise levels in Apple Watch recordings. Indeed, home-based ECG recordings typically suffer from higher noise levels that differ by device and recording method³⁴. In our study, approximately 6% of Apple Watch recordings and 13% of QOCA ECG102D recordings were initially excluded due to poor signal quality, highlighting substantial challenges in obtaining reliable home-based

ECG data. The lower rate of poor-quality recordings with the Apple Watch likely arises from its built-in real-time signal quality assessment, prompting immediate retakes when excessive noise is detected, whereas the QOCA ECG102D assesses quality only after the recording has been completed. There is a study directly applying Lead-I AI models trained on clinical 12-lead ECG data to Apple Watch ECG recordings to predict low EF and reported high accuracy even before device-specific fine-tuning³⁰. However, the accuracy in the other clinical tasks was not reported, which may reflect more substantial accuracy degradation in those tasks. In our study, base models maintained high performance (AUC > 0.8) on Apple Watch data in predicting 90-day mortality, low EF, and elevated NT-proBNP. This device-dependent accuracy could explain the scarcity of published research

Table 3 | Model performance and 95% confidence interval of models on clinical variables stratified by underlying rhythm

	Apple Watch		QOCA ECG102D	
	Sinus rhythm	Atrial fibrillation	Sinus rhythm	Atrial fibrillation
AUC				
Death in 90 days	0.82 (0.73 to 0.91)	0.82 (0.72 to 0.92)	0.75 (0.68 to 0.82)	0.74 (0.63 to 0.85)
EF $\leq$ 40%	0.85 (0.79 to 0.92)	0.71 (0.59 to 0.83)	0.84 (0.80 to 0.89)	0.83 (0.75 to 0.91)
PASP $\geq$ 50 mmHg	0.66 (0.60 to 0.72)	0.55 (0.45 to 0.65)	0.63 (0.58 to 0.68)	0.52 (0.43 to 0.61)
LAD $\geq$ 50 mm	0.76 (0.70 to 0.83)	0.59 (0.49 to 0.69)	0.73 (0.68 to 0.78)	0.60 (0.52 to 0.67)
NT-proBNP $\geq$ 1000 pg/mL	0.83 (0.74 to 0.92)	0.87 (0.69 to 1.00)	0.75 (0.69 to 0.81)	0.75 (0.63 to 0.86)
Hb $\leq$ 10 g/dL	0.74 (0.69 to 0.80)	0.47 (0.27 to 0.66)	0.68 (0.65 to 0.71)	0.68 (0.59 to 0.78)
eGFR $\leq$ 60 mL/min	0.75 (0.71 to 0.78)	0.61 (0.49 to 0.74)	0.71 (0.69 to 0.73)	0.68 (0.62 to 0.74)
Pearson's <i>r</i>				
EF (%)	0.38 (0.34 to 0.41)	0.44 (0.30 to 0.56)	0.35 (0.32 to 0.39)	0.44 (0.33 to 0.53)
PASP (mmHg)	0.14 (0.09 to 0.18)	0.09 (-0.07 to 0.24)	0.18 (0.14 to 0.22)	0.00 (-0.12 to 0.12)
LAD (mm)	0.40 (0.37 to 0.44)	0.35 (0.20 to 0.48)	0.30 (0.27 to 0.34)	0.31 (0.20 to 0.42)
NT-proBNP (pg/mL, log scale)	0.62 (0.48 to 0.74)	0.50 (0.14 to 0.75)	0.48 (0.40 to 0.56)	0.43 (0.22 to 0.60)
Hb (g/dL)	0.43 (0.35 to 0.51)	0.06 (-0.27 to 0.36)	0.33 (0.30 to 0.36)	0.29 (0.16 to 0.42)
eGFR (mL/min)	0.38 (0.33 to 0.44)	0.23 (0.01 to 0.43)	0.35 (0.33 to 0.38)	0.32 (0.22 to 0.41)
MAE				
EF (%)	6.1 (5.9 to 6.4)	8.2 (7.1 to 9.2)	5.9 (5.7 to 6.1)	9.0 (8.1 to 9.9)
PASP (mmHg)	7.0 (6.8 to 7.3)	12.2 (10.4 to 14.0)	7.1 (6.8 to 7.3)	10.5 (9.4 to 11.6)
LAD (mm)	4.5 (4.3 to 4.6)	7.0 (6.0 to 7.9)	4.8 (4.6 to 4.9)	7.3 (6.6 to 8.1)
NT-proBNP (pg/mL, log scale)	0.57 (0.48 to 0.66)	0.48 (0.30 to 0.65)	0.61 (0.56 to 0.66)	0.61 (0.49 to 0.73)
Hb (g/dL)	1.7 (1.6 to 1.8)	2.0 (1.6 to 2.4)	1.5 (1.4 to 1.5)	1.8 (1.6 to 2.0)
eGFR (mL/min)	25.4 (24.1 to 26.7)	27.9 (23.2 to 32.6)	20.5 (20.0 to 21.0)	25.3 (23.2 to 27.4)

AUC area under the receiver operating characteristic curve, MAE mean absolute error, EF ejection fraction, PASP pulmonary artery systolic pressure, LAD left atrial diameter, NT-proBNP N-terminal pro B-type natriuretic peptide, Hb hemoglobin, eGFR estimated glomerular filtration rate.

reporting successful cross-device AI applications beyond these specific outcomes, as direct application of most 12-lead ECG-based AI models to home-based devices often yields unsatisfactory results.

Although direct application of Lead-I models trained on 12-lead ECGs to home-based devices was generally suboptimal, fine-tuning effectively restored accuracy to clinically acceptable levels for most prediction tasks. Specifically, using a performance threshold defined as the external validation AUC minus 0.05, all tasks in the Apple Watch dataset regained this level of accuracy after fine-tuning. In contrast, for the QOCA ECG102D cohort, several tasks—including prediction of 90-day mortality, elevated PASP, elevated NT-proBNP, anemia, and reduced renal function—did not fully return to this threshold, despite significant improvements compared to their baseline performance before fine-tuning. This discrepancy between devices likely reflects a lower initial baseline performance for the QOCA ECG102D, which may stem from greater differences in ECG waveform characteristics compared with Philips ECG signals. Thus, although the QOCA ECG102D cohort in our study had more samples, further data collection and fine-tuning could potentially enhance prediction accuracy even further. Additionally, our analyses validated previous literature demonstrating that Lead-I models trained and validated within 12-lead ECG datasets possess high predictive accuracy^{17,19,20,22–24,26,27}. Although these results cannot directly translate to cross-device scenarios, they offer valuable reference points for understanding which tasks could be effectively reproduced in home-based ECG devices given sufficient fine-tuning data. Our findings demonstrate that an insufficient number of fine-tuning samples, particularly fewer than 1000, can impair performance, occasionally yielding predictions inferior to those of the original base models. This limitation helps explain why previous published studies applying clinical ECG models to home-based devices often avoided fine-tuning, as labeled datasets in these studies typically remained small^{30–33}. Thus, we emphasize the need to encourage future

investigators to initially explore ECG phenotypes extensively using clinical 12-lead datasets—akin to ECG-based phenome-wide association studies⁴⁴—while simultaneously advocating the establishment of large, device-specific cohorts (over 10,000 labeled samples) to ensure practical implementation of home-based ECG applications.

To contextualize performance across the endpoints of this study, we curated benchmark comparisons from prior reduced-lead and wearable ECG literature whenever comparable endpoints and evaluation metrics were available (Table 5). These cross-study comparisons should be interpreted cautiously because published studies differ in phenotype definitions and severity thresholds, prediction horizons, model inputs, and label ascertainment. For example, pulmonary hypertension definitions in prior work¹⁹ may use composite echocardiographic criteria that are not identical to our PASP $\geq$ 50 mmHg threshold; chronic kidney disease (CKD) screening studies may target any-stage CKD²⁰ rather than eGFR $\leq$ 60 mL/min (stage $\geq$ 3); and mortality models may focus on long-term outcomes¹⁷ rather than 90-day death. In addition, some single-lead “ECG” models incorporate demographic covariates (e.g., age and gender) alongside the ECG signal²⁴, and some large-scale wearable studies rely on self-reported disease labels rather than electronic health records (EHR)-linked ground truth, both of which can materially affect the apparent AUC and MAE⁴⁵. Despite these differences, available benchmarks indicate that our Lead-I performance within clinical 12-lead-derived signals is broadly consistent with prior reduced-lead studies^{17–21}, and that our Apple Watch results fall within the range of previously reported smartwatch evaluations⁴⁵. Notably, we found few wearable benchmarks beyond low EF and basic demographics, underscoring a major gap in standardized evaluation resources. To facilitate future benchmarking under unified preprocessing and phenotype definitions, we therefore release HOME dataset as a benchmark.

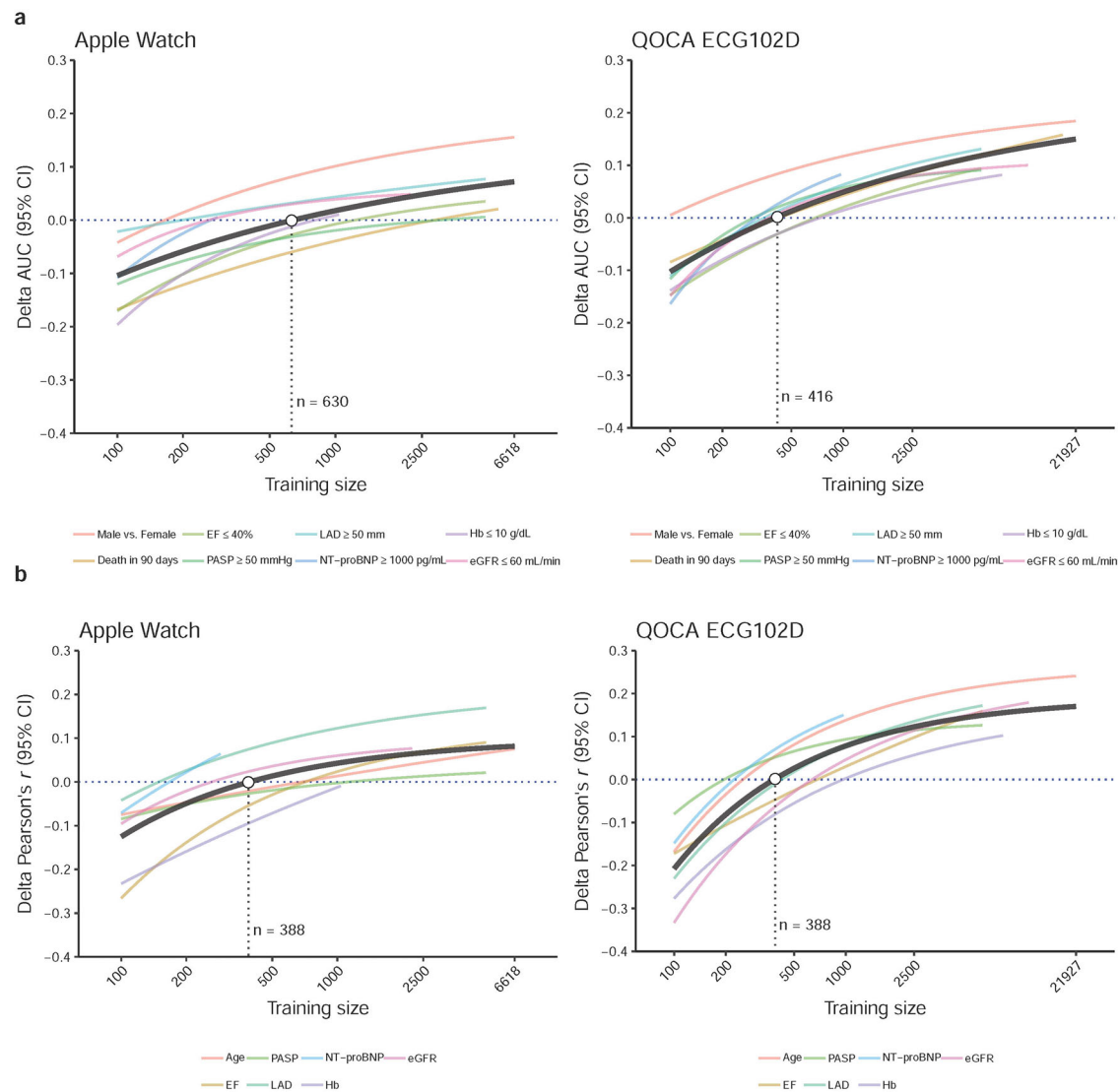


Fig. 5 | Fine-tuning sample size and performance change in home-based ECG models. a The change in discrimination (Delta AUC) for eight binary clinical tasks. **b** The change in continuous prediction accuracy (Delta Pearson's r) for seven continuous variables under the same fine-tuning sample-size scheme. Left panel is for the Apple Watch cohort, right panel for the QOCA ECG102D cohort. The gray line represents the overall learning curve, and colored lines denote individual tasks

(see legend). The horizontal dotted line at zero indicates parity with the base 12-lead model, and open circles with vertical dotted lines mark the approximate sample sizes at which the mean $\Delta AUC/\Delta r$ crosses zero. The number of training sizes is based on the number of ECGs rather than the number of patients. Since the number of patients varies for each task under different training sizes, specific patient numbers are presented in Figures S9 and S10.

Our fine-tuned AI model demonstrated robust predictive performance on Apple Watch ECG data for predicting 90-day mortality, low EF, enlarged LAD, and elevated NT-proBNP, and on QOCA ECG102D data for predicting low EF. Among these, the capability of consumer-grade ECG devices to detect low EF holds promise for opportunistic screening in community settings. Given the widespread adoption of ECG-capable consumer wearables, such approaches could identify previously unrecognized individuals at risk of heart failure⁴⁶, prompting timely referrals for further evaluation and early therapeutic intervention. Prediction of enlarged LAD from single-lead ECGs also presents intriguing clinical possibilities. Enlarged LAD is a known risk factor for AF⁴⁷, and the high predictive accuracy observed, especially during sinus rhythm recordings, suggests these models could help identify patients at increased risk of AF development before clinical manifestation occurs. Additionally, accurate prediction of 90-day mortality has important clinical implications. Prior hospital-based studies have already demonstrated that proactive clinical interventions for patients identified as high risk by AI-enabled ECG models can significantly reduce mortality¹². Extending such predictive capabilities into the home setting could

substantially enhance their benefits by enabling early risk identification and facilitating timely follow-up for individuals otherwise unaware of their elevated risk. The exceptional performance in predicting NT-proBNP levels was also noted, with the model exhibiting not only high classification accuracy but also remarkable correlation with actual NT-proBNP values. Given that NT-proBNP is a well-established biomarker for guiding heart failure management and monitoring therapeutic responses⁴⁸, home-based AI-ECG models capable of accurately predicting NT-proBNP levels could substantially enhance remote management of heart failure patients, potentially reducing hospital readmissions and improving quality of life. However, all continuous prediction tasks exhibited systematic proportional bias in Bland-Altman analyses, consistent with regression-to-the-mean. These characteristics suggest that consumer-grade single-lead ECG models may be more appropriate for opportunistic screening or triage rather than reporting continuous physiologic parameters as standalone quantitative measurements. From a clinical application perspective, our findings support a confirmatory-testing workflow rather than a diagnostic-replacement workflow: wearable AI-ECG can help identify individuals who may benefit

Table 4 | Brief summary of HOME dataset in each task

Device type and task	Total sample size	Number of cases	Mean age	Male proportion	AUC/r/MAE (Lead-I model)	AUC/r/MAE (Fine-tuned model)
Apple Watch						
Binary prediction task						
Gender	1000	N/A	62.5	52.9%	0.71	0.78
Death in 90 days	824	22	63.9	52.4%	0.78	0.79
EF ≤ 40%	761	24	64.0	53.4%	0.74	0.84
PASP ≥ 50 mmHg	762	49	64.0	53.4%	0.73	0.77
LAD ≥ 50 mm	762	35	64.0	53.4%	0.74	0.82
NT-proBNP ≥ 1000 pg/mL	100	38	68.4	50.0%	0.81	0.82
Hb ≤ 10 g/dL	172	38	63.3	55.2%	0.65	0.70
eGFR ≤ 60 mL/min	383	108	64.7	55.1%	0.69	0.74
Continuous prediction task						
Age (years)	1000	N/A	62.5	52.9%	$r = 0.48$ MAE = 11.4	$r = 0.59$ MAE = 10.4
QOCA ECG102D						
Binary prediction task						
Gender	1000	N/A	60.0	52.5%	0.62	0.79
Death in 90 days	823	64	61.4	52.7%	0.50	0.75
EF ≤ 40%	343	29	65.9	53.9%	0.65	0.81
PASP ≥ 50 mmHg	341	29	65.9	53.4%	0.51	0.65
LAD ≥ 50 mm	340	32	65.8	53.8%	0.66	0.68
NT-proBNP ≥ 1000 pg/mL	308	80	64.8	51.9%	0.72	0.82
Hb ≤ 10 g/dL	471	67	59.3	55.8%	0.62	0.72
eGFR ≤ 60 mL/min	655	155	60.6	53.4%	0.64	0.72
Continuous prediction task						
Age (years)	1000	N/A	60.0	52.5%	$r = 0.36$ MAE = 13.0	$r = 0.61$ MAE = 11.0

AUC area under the receiver operating characteristic curve, MAE mean absolute error, EF ejection fraction, PASP pulmonary artery systolic pressure, LAD left atrial diameter, Hb hemoglobin, eGFR estimated glomerular filtration rate.

from targeted echocardiography, laboratory testing, or clinician review, but prospective validation and local calibration are required before routine deployment.

As anticipated, our results confirmed higher predictive accuracy during sinus rhythm compared to AF, a finding consistent with prior research demonstrating poorer AI model performance in patients with comorbidities^{49,50}. However, this rhythm-dependent performance pattern is particularly valuable in consumer-grade ECG devices, as they are initially designed to screen for AF and thus naturally direct AF-positive users toward clinical evaluation^{3–5}. High predictive accuracy in sinus rhythm indicates that, for most users without AF, these AI models could markedly expand the clinical utility of home-based ECG monitoring. Moreover, although prediction performance is generally lower during AF than during sinus rhythm, restricting fine-tuning to sinus rhythm ECGs did not systematically improve sinus rhythm performance. This suggests that rhythm heterogeneity may act as a form of physiological “augmentation” and preserve informative, higher-risk examples that benefit downstream phenotype prediction. Detailed protocols should be developed to guide clinical actions following abnormal home-based ECG predictions, ensuring optimal integration of AI-enabled ECG devices into routine healthcare practice.

This study has several important limitations. First, our analysis was retrospective and relied on ECG recordings and clinical labels integrated into the EHR. Although the single-lead ECGs were patient-acquired outside the conventional ECG laboratory, they were collected in a healthcare-integrated setting and may not fully capture the broader spectrum of motion artifacts and usability variability encountered during unsupervised everyday home monitoring. Consequently, our reported performance should be interpreted as applicable to device-accepted recordings within this

healthcare-integrated workflow, and prospective validation in free-living home environments remains necessary. The prior-to-posterior risk estimates reported here are also empirical test-set estimates under the observed cohort prevalence and should not be interpreted as fully calibrated individual-level probabilities because of the retrospective nature of the study. Second, our analysis included ECG recordings collected predominantly in outpatient or home-based settings rather than acute-care environments such as emergency departments. Consequently, conditions requiring immediate clinical attention, such as severe hyperkalemia or hypokalemia, could not be adequately evaluated in this study. While prior literature has demonstrated the feasibility of predicting dyskalemia using AI models trained exclusively on Lead-I segments³¹, our study design did not allow verification of these findings. Third, although we systematically examined the effect of fine-tuning sample sizes, we could not comprehensively evaluate all potential sample-size thresholds or specific fine-tuning strategies. Specifically, the optimal training approach or hyperparameter settings might differ subtly for each prediction task, but our available computational resources were insufficient to fully explore these variations for every task. Fourth, the wearable-device cohorts reflect patients whose consumer ECG recordings were incorporated into the EHR, which likely selects for individuals with higher health engagement and digital literacy. Although age distributions were similar across cohorts, wearable cohorts exhibited lower prevalence of severe phenotypes and lower short-term mortality, suggesting that frail elderly patients and advanced heart failure populations may be under-represented. Fifth, label sparsity is an inherent limitation of EHR-based retrospective labeling. Several outcomes were available only for very small subsets of ECG recordings (e.g., NT-proBNP in the Apple Watch cohort). As shown in our results, predictive performance

Table 5 | Comparative performance of Lead-I AI models between the present study and prior studies

	Lead-I model performance in Lead-I segment from 12-lead ECGs		Lead-I model performance in Apple Watch ECGs		Apple Watch fine-tuning model performance in Apple Watch ECGs			
	This study		Prior studies		This study	Prior studies	This study	Prior studies
Gender	AUC = 0.83 (internal) AUC = 0.82 (external)		AUC = 0.80 (internal) ²¹				AUC = 0.78	^a AUC = 0.85 ⁴⁵
Death in 90 days	AUC = 0.87 (internal) AUC = 0.86 (external)		^a AUC = 0.75 (internal) ¹⁷					
EF ≤ 40%	AUC = 0.88 (internal) AUC = 0.88 (external)		AUC = 0.87 (internal) ¹⁸ AUC = 0.93 (external) ¹⁸		AUC = 0.81	^d AUC = 0.88 ³⁰	AUC = 0.85	^a AUC = 0.84 ⁴⁵
PASP ≥ 50 mmHg	AUC = 0.77 (internal) AUC = 0.78 (external)		^b AUC = 0.82 (internal) ¹⁹ ^d AUC = 0.82 (external) ¹⁹					
eGFR ≤ 60 mL/min	AUC = 0.81 (internal) AUC = 0.79 (external)		^c AUC = 0.74 (internal) ²⁰				AUC = 0.75	^a AUC = 0.67 ⁴⁵
Age	MAE = 11.4 (internal) MAE = 12.0 (external)		MAE = 11.6 (internal) ²¹				MAE = 10.5	^a MAE = 9.0 ⁴⁵

AUC area under the receiver operating characteristic curve, MAE mean absolute error, EF ejection fraction, PASP pulmonary artery systolic pressure, eGFR estimated glomerular filtration rate.

^aIn the study by Sau et al., the output of the Lead-I model was 20-year mortality¹⁷.

^bIn the study by Kwon et al., pulmonary hypertension was defined as a tricuspid regurgitation peak gradient >32 mm Hg, an estimated PASP > 35 mm Hg, or a tricuspid regurgitation peak gradient ≤32 mm Hg in the presence of at least two additional echocardiographic signs (tricuspid annular plane systolic excursion <16 mm, right ventricular fractional area change <35%, tricuspid annular systolic velocity <10 cm/s, or right ventricular dilatation)¹⁹.

^cIn the study by Holmström et al., the output of the Lead-I model was the presence of chronic kidney disease at any stage¹⁷.

^dIn the study by Attia et al., a retrained single-lead AI algorithm was reported to be used to analyze Apple Watch ECGs for predicting an ejection fraction ≤40%; however, the methodological details were not disclosed, making it unclear whether the model was specifically fine-tuned using Apple Watch ECG data²⁰.

^aAbbaspourazad et al. analyzed more than 3.7 million Apple Watch ECG recordings, of which approximately 17,000 were linked to self-reported disease status⁴⁵.

continues to improve with increasing sample size, suggesting that the accuracy reported in this study should not be regarded as the upper bound of achievable performance. This interpretation is supported by prior large-scale work involving more than 3.7 million Apple Watch ECGs, which demonstrated superior accuracy for gender and age prediction compared with the present study⁴⁵. Finally, our results were derived from a single ethnic group and geographically proximate hospitals. In addition, no independent external wearable ECG cohort was available; Apple Watch and QOCA ECG102D performance should therefore be interpreted as device-specific internal validation. Future research should incorporate multi-ethnic data to ensure broad generalizability and practical applicability of home-based AI-ECG models.

In conclusion, our study systematically quantified the accuracy loss associated with cross-device domain shifts when applying AI-enabled Lead-I ECG models—trained on clinical-grade 12-lead ECG data—to consumer-grade home-based ECG devices. Although significant accuracy declines were observed with direct cross-device model application, targeted fine-tuning using device-specific ECG data largely restored predictive accuracy across multiple clinically important outcomes, particularly for detecting low EF, predicting 90-day mortality, identifying enlarged LAD, and estimating elevated NT-proBNP. These findings highlight both the necessity and feasibility of fine-tuning for practical implementation of AI models in home-based ECG monitoring. Future efforts should focus on prospective validation studies conducted in diverse populations and real-world home environments to confirm clinical utility, ultimately facilitating the integration of AI-ECG applications into routine healthcare practice and improving preventive healthcare strategies.

Methods

Data source and ethical statement

This study was approved by the Institutional Review Board (IRB) of Tri-Service General Hospital (A202505026), which also granted a waiver of informed consent. The waiver was justified because this was a retrospective analysis of ECG recordings that had been collected by patients using consumer-grade devices—specifically, the Apple Watch (Taiwan Food and Drug Administration [TFDA] approval no. 033901) and the QOCA ECG102D (TFDA approval no. 005325)—and routinely integrated into patients' electronic health records (EHRs) as part of standard care. Given the retrospective design and the use of existing EHR-linked data, the requirement for individual patient informed consent was waived by the IRB. Importantly, these single-lead ECGs represent patient-initiated recordings integrated into routine care via the EHR; we did not analyze continuous free-

living recordings captured during daily activities. Therefore, generalizability to fully unsupervised home monitoring requires prospective evaluation.

The workflow and dataset details are summarized in Fig. 1. Briefly, we developed a base AI model using Lead-I ECG segments extracted from standard clinical 12-lead ECG recordings from the Philips TC series (TC30, TC50, TC70) acquired at the Neihu General Hospital site of Tri-Service General Hospital (academic medical center). A total of 246,874 patients were randomly assigned to a training set (172,827 patients, 474,130 ECGs) and an internal test set (74,047 patients, 202,062 ECGs). Additionally, an independent external test set comprising 65,338 patients (219,231 ECGs) from Tingzhou Branch Hospital (community hospital) was used for external validation.

To evaluate the generalizability of the base AI model to home-based ECG recordings, we collected single-lead ECG data from two types of consumer-grade devices: Apple Watch and QOCA ECG102D. For both consumer devices, we used the vendor-provided signal-quality assessment as the first-line quality gate and excluded ECGs flagged as “inconclusive”, “poor recording”, or “low quality” prior to model training and evaluation. We only included the “sinus rhythm” and “atrial fibrillation” recordings identified by vendor algorithms. This choice reflects real-world operation of these devices, where low-quality recordings are typically rejected or prompt reacquisition, and thus our AI models are intended to be applied to device-accepted recordings integrated into the EHR. Specifically, for the Apple Watch dataset, 8790 patients at Neihu General Hospital recorded Lead-I ECGs. Patients were randomly assigned to training (6113 patients, 7044 ECGs) and test (2677 patients, 3399 ECGs) sets. After excluding ECG recordings classified by the device vendor as inconclusive or poor quality (426 ECGs [6.0%] in the training set and 182 ECGs [5.4%] in the test set), we retained 6618 ECGs in the training set and 3217 ECGs in the test set. Similarly, the QOCA ECG102D dataset included 27,983 patients who recorded ECGs. After randomization into training (19,553 patients, 24,977 ECGs) and test (8430 patients, 10,942 ECGs) subsets, ECG signals labeled as low quality by the device vendor (3050 ECGs [13.9%] in the training set and 1352 ECGs [12.4%] in the test set) were excluded, resulting in final training and test sets comprising 21,927 ECGs and 9590 ECGs, respectively.

All 12-lead ECGs including recordings from both the Neihu and Tingzhou sites, were obtained before February 2022. In contrast, the Apple Watch and QOCA ECG102D cohorts were composed predominantly of recordings obtained after January 2023. Because these data were collected as part of longitudinal routine clinical care, there was partial patient overlap across cohorts. Specifically, 54% of patients in the QOCA ECG102D cohort and 58% of patients in the Apple Watch cohort had prior 12-lead ECG data

in Cohort A. In addition, 15% of patients in the QOCA ECG102D cohort also appeared in the Apple Watch cohort, and 47% of patients in the Apple Watch cohort also appeared in the QOCA ECG102D cohort. To prevent patient-level leakage despite partial overlap across cohorts, patients were not randomized independently within each cohort. Instead, we generated a single master patient-level assignment from the union of all eligible medical record numbers across Cohorts A–C and allocated these medical record numbers to the development or test partition at a 70%/30% ratio. This same assignment was then applied to every cohort. Thus, any patient appearing in more than one cohort was assigned to the same partition across all cohorts; for example, a patient in the development set of Cohort A could only appear in the development set of the Apple Watch or QOCA ECG102D cohorts, and could not appear in their test sets. This design prevented patient-level information leakage between development and test partitions despite cross-cohort overlap.

ECG signal preprocessing

The original ECG format obtained from the Philips TC-series system was a 10-s, 12-lead recording, whereas ECG recordings from the Apple Watch and QOCA ECG102D were 30-s single-lead (Lead-I) recordings. To ensure consistency across datasets, we extracted Lead-I segment from the 12-lead ECG recordings and standardized the amplitude unit to 10 μ V increments for all three ECG sources. Additionally, we uniformly resampled signals to 500 Hz, as the original sampling frequency was 512 Hz for Apple Watch data, whereas Philips TC-series and QOCA ECG102D data were sampled at 500 Hz.

Regarding filtering, the 12-lead ECGs from the Philips system had already undergone built-in filtering with a high-pass filter at 0.05 Hz and a low-pass filter at 150 Hz. In contrast, despite initial vendor-applied filtering of the Apple Watch and QOCA ECG102D signals, significant baseline wander and noise persisted. Therefore, we implemented additional digital filtering for these two datasets using second-order Butterworth high-pass (0.5 Hz) and low-pass (150 Hz) filters to effectively reduce baseline drift and high-frequency noise. To ensure consistency, given that the Apple Watch allows user recordings shorter than 30 s, we uniformly trimmed Apple Watch and QOCA ECG102D recordings to retain only the central 28-s segments. For the Philips TC-series ECG recordings, we utilized the entire 10-s duration without additional truncation.

Phenotypes used for prediction

Since each ECG recording had a precise timestamp logged within the EHR, we utilized this timestamp as the index time to merge the ECG data with patient demographic and clinical variables, echocardiographic measurements, and laboratory test results. Initially, we performed a literature review to identify appropriate phenotypes for prediction. The phenotypes selected had to meet two criteria: (1) previous studies had demonstrated the feasibility of predicting these phenotypes from Lead-I ECG recordings; (2) sufficient abnormal samples were available within our dataset. Based on these criteria, we identified a total of 9 phenotypes for prediction, categorized as follows: Demographic and clinical variables—gender, age, 90-day mortality; Echocardiographic outcomes—left ventricular ejection fraction (EF), pulmonary artery systolic pressure (PASP), left atrial diameter (LAD); Laboratory outcomes—N-terminal pro-B-type natriuretic peptide (NT-proBNP), hemoglobin (Hb), estimated glomerular filtration rate (eGFR). Gender and Age predictions were retained as a methodological control and domain-shift indicator rather than a clinically actionable target. We merged echocardiographic outcomes within ± 7 days of the index ECG recording time³⁶. Laboratory measurements were merged using different time windows based on their clinical relevance and frequency of testing: NT-proBNP within ± 24 h³⁷, Hb within ± 12 h³⁸, and eGFR within ± 30 days³⁹.

From these phenotypes, we further defined eight binary prediction tasks—Gender (male vs. female), death in 90 days (death vs. survival), low EF ($\leq 40\%$ vs. $>40\%$), elevated PASP (≥ 50 mmHg vs. <50 mmHg), enlarged LAD (≥ 50 mm vs. <50 mm), elevated NT-proBNP (≥ 1000 pg/mL vs. <1000 pg/mL), anemia (Hb ≤ 10 g/dL vs. >10 g/dL), reduced renal function

(eGFR ≤ 60 mL/min vs. >60 mL/min) –and seven continuous prediction tasks—age (years), EF (%), PASP (mmHg), LAD (mm), NT-proBNP (pg/mL) in log-transformed (log base 10) due to severe right skewness, Hb (g/dL), eGFR (mL/min).

Deep learning model training and inference

The deep learning architecture utilized in this study was based on the ResNet architecture, employing one-dimensional convolutional layers for feature extraction. The detailed network architecture has been previously described⁵¹. The only modification was the removal of the original module designed for merging 12-lead ECG information. High-level features extracted directly from Lead-I ECG segments were instead used for subsequent binary or continuous predictions. All models were trained using the Adam optimizer (standard parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$), with an initial learning rate of 0.0001, reduced by a factor of 10 at the 30th and 60th epochs, a mini-batch size of 32 ECG segments, and weight decay of 0.001. Training was performed for a total of 100 epochs. In each task, the training data were randomly partitioned, allocating 80% for model fitting and 20% to select the best-performing model based on lowest validation loss. For each prediction task, only a single optimal model was selected for final inference on the test set.

For data augmentation, random cropping and signal inversion (vertical flipping) were applied. Specifically, during training with Lead-I segments from 12-lead ECGs (original length: 5000 data points), random crops of 4096 points were used. At inference, predictions from four signal variations (original first and last 4096 points, and vertically flipped first and last 4096 points) were averaged. For Apple Watch and QOCA ECG102D signals (original length after preprocessing: 14,000 data points), similar random cropping to a length of 12,288 data points was performed. Inference similarly averaged predictions from four variations (original first and last 12,288 points and vertically flipped first and last 12,288 points). Due to the use of global average pooling in the network's final layer, the model initially trained on 4096 data points was fully compatible with longer input lengths, including 12,288 data points, allowing direct application to Apple Watch and QOCA ECG102D data without further adjustment.

The initial weights for models trained on Lead-I segments derived from the 12-lead ECG dataset were randomly initialized. Conversely, all fine-tuned models employed weights from the corresponding 12-lead ECG-trained Lead-I model as initial parameters. Fine-tuning experiments utilized subsets of varying sample sizes ($n = 100, 200, 500, 1000$, and 2500 ECGs). For subsets with particularly small training samples, we ensured inclusion of at least two positive cases. For binary prediction tasks, oversampling was implemented to ensure balanced representation, enabling the model to learn discriminative features from both case and control samples. For continuous prediction tasks, the range of values from the minimum to the maximum was partitioned into 20 equal bins, and samples within each bin were oversampled to achieve equal counts⁵¹. Because this strategy alters the value distribution of the training set, we subsequently applied a linear regression-based calibration to adjust the predicted values and mitigate potential numerical offset. Moreover, we also performed a sensitivity analysis using an inverse regression calibration to address regression-to-the-mean. All calibration mappings were derived exclusively from the validating subset and then applied unchanged to the independent test set.

Statistical analysis and model performance assessment

Continuous variables were presented as means $\pm$ standard deviation, while categorical variables were summarized as counts and percentages. The performance of AI models for binary classification tasks was evaluated using the AUC along with corresponding 95% confidence intervals (CI). For continuous prediction tasks, the performance was quantified using Pearson's correlation coefficient (r) and MAE with associated 95% CIs. To rigorously assess domain shifts from 12-lead ECG-trained models to home-based single-lead ECG devices, we compared the prediction performance between the original Lead-I model trained on 12-lead ECGs and fine-tuned models using data collected from Apple Watch or QOCA ECG102D

devices. Model performance comparisons were tested using DeLong's method for differences in AUC, Fisher's Z-transformation for differences in Pearson's correlation coefficients, and paired *t*-test for difference in MAE. To assess agreement and potential systematic bias for continuous endpoints, we additionally performed Bland–Altman analyses, plotting prediction error (AI-ECG prediction minus ground truth) against the corresponding ground-truth measurement; the mean difference (bias) and 95% limits of agreement were defined as mean $\pm$ 1.96 standard deviations.

For 7 clinical phenotypes, we also reported sensitivity and specificity at a pre-specified operating threshold. For clinical interpretability, we additionally reported prior risk as the observed test-set prevalence, posterior risk after a positive AI-ECG screen as positive predictive value (PPV), and residual risk after a negative AI-ECG screen as 1 - negative predictive value (NPV). Two thresholds were evaluated. First, a Youden-index threshold was selected on the validating subset (20% split from the training set) to provide a balanced operating point for model comparison. Second, a sensitivity-prioritized threshold was selected on the validation subset to target 90% sensitivity, reflecting a screening-oriented workflow in which missed cases may be more harmful than additional confirmatory testing. Both thresholds were selected exclusively in the validation subset and then applied unchanged to the independent test set. Accordingly, the realized sensitivity in the test set, particularly in small subgroups, may differ from the nominal validation target.

Additionally, to evaluate the relationship between the fine-tuning sample size and improvement in prediction performance, we generated learning curves showing changes in AUC (binary tasks) and Pearson's *r* (continuous tasks) relative to increasing fine-tuning dataset sizes. The relationship between sample size and predictive accuracy was modeled using a power-law function according to a previous study⁵². This analysis allowed us to quantify the minimal number of fine-tuning samples required to restore performance lost due to insufficient fine-tuning samples. All statistical analyses were performed using R software (version 3.4.4), and two-tailed *p*-values < 0.05 were considered statistically significant.

Benchmark dataset generation

We curated and released an evaluation benchmark, the Health records-linked Open Multi-consumer device Electrocardiogram (HOME) dataset, from the Apple Watch test set and the QOCA ECG102D test set. For each device, we constructed a patient-level subset of 1000 individuals, with exactly one single-lead ECG recording per patient to avoid within-subject correlation. We excluded the Apple Watch ECGs shorter than 30 s. Because 90-day mortality events and the availability of linked NT-proBNP measurements were sparse in the wearable test cohorts, we used an enriched two-stage sampling strategy to ensure adequate representation of these endpoints. Specifically, within each device-specific test cohort, we first included all eligible patients who (i) died within 90 days of the index ECG and/or (ii) had an NT-proBNP value available within the prespecified linkage window. If multiple eligible ECGs existed for a patient, one recording was randomly selected. After forcing inclusion of the above patients, the remaining slots were filled by simple random sampling from the rest of the device-specific test cohort until a total of 1000 patients was reached.

To mitigate privacy risks, the released HOME waveforms were downsampled to 200 Hz, resulting in 6000 samples per 30-s recording. For compatibility with the original inference pipeline (trained and evaluated using 500 Hz signals), we upsampled the released waveforms back to 15,000 samples (corresponding to 30 s at 500 Hz) and applied the same preprocessing and inference procedure as in the main analyses. Task-specific model predictions (before and after device-specific fine-tuning) are released alongside the waveform files to facilitate standardized benchmarking and reproducibility.

Data availability

To support transparency and reproducibility while protecting patient privacy and complying with institutional and regulatory requirements, this

study adopts a tiered data-access strategy. An evaluation-only version of the Health records-linked Open Multi-consumer device Electrocardiogram (HOME) dataset is available on GitHub (<https://github.com/xup6fup/HOME>) and has been archived in Zenodo (<https://doi.org/10.5281/zenodo.18014346>)⁵³, consisting of single-lead ECG waveforms from consumer-grade devices without ground-truth labels. This benchmark is designed exclusively for inference-based model evaluation. Users may submit task-specific predictions for centralized performance assessment, but ground-truth labels are intentionally withheld to prevent data leakage, implicit domain adaptation, and post-hoc calibration. The full raw ECG datasets and label-linked clinical data are not publicly released because they are derived from electronic health records and are subject to institutional and regulatory restrictions. Researchers with legitimate scientific interests may request access through a formal application process to authors (C. Lin, xup6fup0629@gmail.com; C.S. Lin, littlelincs@gmail.com), which requires approval by the Tri-Service General Hospital Institutional Review Board and execution of a research-use data agreement. Approved researchers will be granted remote access under a secure VPN environment to a designated server hosted by Tri-Service General Hospital, where model development and analysis may be conducted. Only aggregate summary outputs, such as tables, figures, and statistical results, may be retrieved following review and approval.

Code availability

Training and evaluation scripts used in this study are publicly available on GitHub (<https://github.com/xup6fup/HOME>) and have been archived in Zenodo (<https://doi.org/10.5281/zenodo.18014346>)⁵³. Predictions from both the base model and the fine-tuned models are provided on the HOME dataset to enable direct comparison. Pretrained model weights and related inference code will be made available for non-commercial academic research upon a reasonable request to the corresponding authors, subject to institutional policies and execution of an appropriate research agreement.

Received: 8 October 2025; Accepted: 12 June 2026;

Published online: 06 July 2026

References

- Grady, P. A. & Gough, L. L. Self-management: a comprehensive approach to management of chronic conditions. *Am. J. Public Health* **104**, e25–e31 (2014).
- Huhn, S. et al. The impact of wearable technologies in health research: scoping review. *JMIR Mhealth Uhealth* **10**, e34384 (2022).
- Ha, A. C. T. et al. Effect of continuous electrocardiogram monitoring on detection of undiagnosed atrial fibrillation after hospitalization for cardiac surgery: a randomized clinical trial. *JAMA Netw. Open* **4**, e2121867 (2021).
- Halcox, J. P. J. et al. Assessment of remote heart rhythm sampling using the AliveCor heart monitor to screen for atrial fibrillation: The REHEARSE-AF study. *Circulation* **136**, 1784–1794 (2017).
- Shahid, S. et al. Diagnostic accuracy of Apple Watch electrocardiogram for atrial fibrillation: a systematic review and meta-analysis. *JACC Adv.* **4**, 101538 (2025).
- Attia, Z. I., Harmon, D. M., Behr, E. R. & Friedman, P. A. Application of artificial intelligence to the electrocardiogram. *Eur. Heart J.* **42**, 4717–4730 (2021).
- Lin, C. S., Liu, W. T., Chen, Y. H., Lin, S. H. & Lin, C. Artificial intelligence-enabled electrocardiography from scientific research to clinical application. *EMBO Mol. Med.* **18**, 22–40 (2025).
- Attia, Z. I. et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat. Med.* **25**, 70–74 (2019).
- 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 (2019).

10. Raghunath, S. et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat. Med.* **26**, 886–891 (2020).
11. Yao, X. et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nat. Med.* **27**, 815–819 (2021).
12. Lin, C. S. et al. AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial. *Nat. Med.* **30**, 1461–1470 (2024).
13. Noseworthy, P. A. et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. *Lancet* **400**, 1206–1212 (2022).
14. Tsai, D. J. et al. Artificial intelligence-assisted diagnosis and prognostication in low ejection fraction using electrocardiograms in inpatient department: a pragmatic randomized controlled trial. *BMC Med.* **23**, 342 (2025).
15. Liu, W. T. et al. Artificial intelligence-enabled ECGs for atrial fibrillation identification and enhanced oral anticoagulant adoption: a pragmatic randomized clinical trial. *J. Am. Heart Assoc.* **14**, e042106 (2025).
16. Lin, C. et al. Artificial intelligence-powered rapid identification of ST-elevation myocardial infarction via electrocardiogram (ARISE)—a pragmatic randomized controlled trial. *NEJM AI* **1**, A10a2400190 (2024).
17. Sau, A. et al. Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: a model development and validation study. *Lancet Digit Health* **6**, e791–e802 (2024).
18. Cho, J. et al. Artificial intelligence algorithm for screening heart failure with reduced ejection fraction using electrocardiography. *ASAIO J.* **67**, 314–321 (2021).
19. Kwon, J. M. et al. Artificial intelligence for early prediction of pulmonary hypertension using electrocardiography. *J. Heart Lung Transpl.* **39**, 805–814 (2020).
20. Holmstrom, L. et al. Deep learning-based electrocardiographic screening for chronic kidney disease. *Commun. Med.* **3**, 73 (2023).
21. Dias, F. M., Ribeiro, E., Soares, Q. B., Krieger, J. E. & Gutierrez, M. A. Predicting age and sex from reduced lead electrocardiograms using deep learning. In *Simpósio Brasileiro de Computação Aplicada à Saúde*. <https://doi.org/10.5753/sbcas.2025.7705> (2025).
22. Dupulthys, S. et al. Single-lead electrocardiogram artificial intelligence model with risk factors detects atrial fibrillation during sinus rhythm. *Europace* **26**, eua4354 (2024).
23. Khunte, A. et al. Detection of left ventricular systolic dysfunction from single-lead electrocardiography adapted for portable and wearable devices. *npj Digit Med.* **6**, 124 (2023).
24. Kwon, J. M. et al. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. *Lancet Digit Health* **2**, e358–e367 (2020).
25. Dhingra, L. S. et al. Artificial intelligence-enabled prediction of heart failure risk from single-lead electrocardiograms. *JAMA Cardiol.* **10**, 574–584 (2025).
26. Lin, C. et al. Point-of-care artificial intelligence-enabled ECG for dyskalemia: a retrospective cohort analysis for accuracy and outcome prediction. *npj Digit Med.* **5**, 8 (2022).
27. Liu, W. C. et al. A deep-learning algorithm for detecting acute myocardial infarction. *EuroIntervention* **17**, 765–773 (2021).
28. Reyna, M. A. et al. Will two do? Varying dimensions in electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021 In *Computing in Cardiology 2021*. 1–4. <https://doi.org/10.23919/CinC53138.2021.9662687>
29. Reyna, M. A. et al. Issues in the automated classification of multilead ECGs using heterogeneous labels and populations. *Physiol. Meas.* **43**, 084001 (2022).
30. Attia, Z. I. et al. Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction. *Nat. Med.* **28**, 2497–2503 (2022).
31. Chiu, I. M. et al. Serum potassium monitoring using AI-enabled smartwatch electrocardiograms. *JACC Clin. Electrophysiol.* **10**, 2644–2654 (2024).
32. Lim, J. et al. Artificial intelligence-enhanced six-lead portable electrocardiogram device for detecting left ventricular systolic dysfunction: a prospective single-centre cohort study. *Eur. Heart J. Digit Health* **6**, 476–485 (2025).
33. Schlesinger, D. E. et al. Artificial intelligence for hemodynamic monitoring with a wearable electrocardiogram monitor. *Commun. Med.* **5**, 4 (2025).
34. Mannhart, D. et al. Clinical validation of 5 direct-to-consumer wearable smart devices to detect atrial fibrillation: BASEL wearable study. *JACC Clin. Electrophysiol.* **9**, 232–242 (2023).
35. Hnatkova, K. et al. Sex differences in heart rate responses to postural provocations. *Int. J. Cardiol.* **297**, 126–134 (2019).
36. Chen, H. Y. et al. Artificial intelligence-enabled electrocardiography predicts left ventricular dysfunction and future cardiovascular outcomes: a retrospective analysis. *J. Pers. Med.* **12**, 455 (2022).
37. Liu, P. Y. et al. Artificial intelligence-enabled electrocardiography detects B-type natriuretic peptide and N-terminal pro-brain natriuretic peptide. *Diagnostics* **13**, 2723 (2023).
38. Yang, C. C. et al. Non-invasive hemoglobin estimation with outcome prediction via deep learning analysis of ECG-derived cardiac micro-dynamics. *Intern. Emerg. Med.* **21**, 525–536 (2026).
39. Tsai, P. H., Lee, S. Y., Wei, C. H., Hsu, Y. J. & Lin, C. ECG-based deep learning for chronic kidney disease detection and cardiovascular risk prediction. *BMC Med. Inf. Decis. Mak.* **25**, 439 (2025).
40. Tsai, D. J. et al. Mortality risk prediction of the electrocardiogram as an informative indicator of cardiovascular diseases. *Digit Health* **9**, 20552076231187247 (2023).
41. Lee, C. H. et al. Artificial intelligence-enabled electrocardiogram screens low left ventricular ejection fraction with a degree of confidence. *Digit Health* **8**, 20552076221143249 (2022).
42. Obata, Y. et al. The impact of posture on the cardiac depolarization and repolarization phases of the QT interval in healthy subjects. *J. Electrocardiol.* **50**, 640–645 (2017).
43. Malik, M., Hnatkova, K., Sisakova, M. & Schmidt, G. Subject-specific heart rate dependency of electrocardiographic QT, PQ, and QRS intervals. *J. Electrocardiol.* **41**, 491–497 (2008).
44. Friedman, S. F. et al. Unsupervised deep learning of electrocardiograms enables scalable human disease profiling. *npj Digit Med.* **8**, 23 (2025).
45. Abbaspourazad, S. et al. Large-scale training of foundation models for wearable biosignals. In *International Conference on Learning Representations (ICLR)* <https://arxiv.org/abs/2312.05409> (2024).
46. McDonagh, T. A., McDonald, K. & Maisel, A. S. Screening for asymptomatic left ventricular dysfunction using B-type natriuretic peptide. *Congest. Heart Fail.* **14**, 5–8 (2008).
47. Vaziri, S. M., Larson, M. G., Benjamin, E. J. & Levy, D. Echocardiographic predictors of nonrheumatic atrial fibrillation. The Framingham Heart Study. *Circulation* **89**, 724–730 (1994).
48. Ponikowski, P. et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur. J. Heart Fail.* **18**, 891–975 (2016).
49. Lou, Y. S. et al. Artificial intelligence-enabled electrocardiogram estimates left atrium enlargement as a predictor of future cardiovascular disease. *J. Pers. Med.* **12**, 315 (2022).
50. Chen, H. Y. et al. Artificial intelligence-enabled electrocardiogram predicted left ventricle diameter as an independent risk factor of long-term cardiovascular outcome in patients with normal ejection fraction. *Front. Med.* **9**, 870523 (2022).

51. Lou, Y. S., Lin, C. S., Fang, W. H., Lee, C. C. & Lin, C. Extensive deep learning model to enhance electrocardiogram application via latent cardiovascular feature extraction from identity identification. *Comput. Methods Prog. Biomed.* **231**, 107359 (2023).
52. Figueroa, R. L., Zeng-Treitler, Q., Kandula, S. & Ngo, L. H. Predicting sample size required for classification performance. *BMC Med. Inf. Decis. Mak.* **12**, 8 (2012).
53. Lin, C. Health records-linked open multi-consumer device electrocardiogram (HOME) dataset. Zenodo <https://doi.org/10.5281/zenodo.18014346> (2025).

Acknowledgements

This study was supported by funding from the National Science and Technology Council, Taiwan (NSTC 114-2321-B-016-005, NSTC 115-2321-B-016-003 to S.H. Lin and NSTC 112-2314-B-016-049, NSTC 112-2410-H-016-002 to C.S. Lin), the Tri-Service General Hospital, Taiwan (TSGH-E-114215, TSGH-C01-113027 to C.S. Lin), and the Cheng Hsin General Hospital, Taiwan (CHNDMC-113-01, CHNDMC-113-11207 to C.S. Lin), and National Defense Medical University, Taiwan (MND-MAB-C02-114005 and MND-MAB-C07-113022 to C.S. Lin).

Author contributions

Chin Lin and Chin-Sheng Lin contributed to the conception and design of the study. Wei-Ting Liu, Da-Wei Chang, Tsung-Neng Tsai, and Cheng-Chung Cheng contributed to the interpretation of the data. Wei-Ting Liu, Da-Wei Chang, Tsung-Neng Tsai, and Cheng-Chung Cheng contributed to the acquisition of data. Chin Lin and Kai-Chieh Chen trained the deep learning models. Chin Lin and Dung-Jang Tsai analyzed the data. Chin Lin drafted the initial manuscript. Wei-Ting Liu, Chih-Yuan Lin, Yuan-Hao Chen, Chien-Sung Tsai, Shih-Hua Lin, and Chin-Sheng Lin revised the manuscript for important intellectual content. Chin-Sheng Lin takes final responsibility for the integrity of the work and approved the final version of the manuscript.

Competing interests

Author Chin Lin is Associate Editor of npj Digital Medicine. Chin Lin was not involved in the journal's review of, or decisions related to, this manuscript. National Defense Medical University intends to license intellectual property related to the AI technology described in this study to Quanta Computer Inc.

Financial benefits arising from such licensing (e.g., royalties) may accrue to Chin Lin, Wei-Ting Liu, Kai-Chieh Chen, Da-Wei Chang, Tsung-Neng Tsai, Cheng-Chung Cheng, Chih-Yuan Lin, Yuan-Hao Chen, Chien-Sung Tsai, Shih-Hua Lin, Chin-Sheng Lin. All other authors have stated they have no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-026-02919-5>.

Correspondence and requests for materials should be addressed to Chin-Sheng Lin.

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.

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

© The Author(s) 2026