Artificial Intelligence-Enhanced Electrocardiogram Models for Detection of Left Ventricular Dysfunction: A Comparison Study

Philip M. Croon, MD[1,2*], Machteld J. Boonstra, PhD[1,3*], Cornelis P. Allaart, MD, PhD[1], Bauke K.O. Arends, MD, Lovedeep S. Dhingra[2], MBBS, MHS, Yu-Chang Huang, MD[5], Thomas Mast, MD, PhD[6], Rohan Khera, MD, MS[2,7,8], Chang-Fu Kuo, MD, PhD[9,10], Joon-Myoung Kwon, MD, MS[11,12], Hak-Seung Lee, MD[11,12], Min Sung Lee, MD[11,12], Rutger R. van de Leur, MD, PhD[4], Zhi-Yong Liu, PhD[13], Evangelos K. Oikonomou, MD, DPhil[2], Jasper L. Selder, MD[1], Michiel M. Winter, MD, PhD[1], Folkert W. Asselbergs, MD, PhD[1,14,15]

1. Department of Cardiology, Amsterdam Cardiovascular Science, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands.
2. Section of Cardiovascular Medicine, Department of Internal Medicine, Yale University, New Haven, USA
3. Department of Medical Informatics, Amsterdam UMC Location University of Amsterdam, Amsterdam, Netherlands.
4. Department of Cardiology, University Medical Center Utrecht, Utrecht, The Netherlands
5. Division of Cardiology, Chang Gung Memorial Hospital, Taoyuan, Taiwan
6. Department of Cardiology, Catharina Hospital Eindhoven, Eindhoven, The Netherlands
7. Center for Outcomes Research and Evaluation, Yale New Haven Hospital, New Haven, CT, USA
8. Section of Health Informatics, Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA
9. School of Medicine, College of Medicine, Chang Gung University, Taoyuan 33302, Taiwan
10. Division of Rheumatology, Allergy and Immunology, Chang Gung Memorial Hospital, Taoyuan, Taiwan.
11. Digital Healthcare Institute, Sejong Medical Research Institute, Bucheon, Republic of Korea

12. Medical AI Co., Ltd., Seoul, Republic of Korea
13. Center for Artificial Intelligence in Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan
14. Institute of Health Informatics, University College London, London, United Kingdom
15. The National Institute for Health Research UCL Hospitals Biomedical Research Centre, University College London, London, United Kingdom

* Equal contribution

Running title: AI-ECG for LVSD

Word count text: 3816

Word count abstract: 251

Number of references: 52

Number of tables and figures: 6

Disclosures: FWA is supported by UCL Hospitals NIHR Biomedical Research Centre. The Department of Cardiology at UMC Utrecht may receive royalties in the future from sales of deep learning ECG algorithms developed by Cordys Analytics, a spin-off company. Additionally, RvE and RvdL are shareholders of Cordys Analytics. EKO reported being a cofounder of Evidence2Health, serving as a consultant to Caristo Diagnostics Ltd and Ensight-AI, having stock options in Caristo Diagnostics Ltd, receiving a grant from the National Heart, Lung, and Blood Institute of the National Institutes of Health, and having patents 63/508,315 and 63/177,117 outside the submitted work. RK reported receiving grants from the National Heart, Lung, and Blood Institute, National Institutes of Health, Doris Duke Charitable Foundation, Bristol Myers Squibb, Novo Nordisk, BridgeBio, and Blavatnik Foundation, being an academic cofounder of Ensight-AI and Evidence2Health, having patents 63/346,610, WO2023230345A1, US20220336048A1, 63/484,426, 63/508,315, 63/580,137, 63/606,203, 63/619,241, and 63/562,335 pending, and serving

as associate editor of JAMA outside the submitted work. No other disclosures were reported. These affiliations and potential financial interests have been disclosed and are being managed in accordance with institutional policies.

Funding: PC is supported by the University of Amsterdam Research Priority Agenda Program AI for Health Decision-Making. This work was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee EP/Z000211/1. This work received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No. 101057849 (DataTools4Heart project) and No. 101080430 (AI4HF project).

Correspondence:

Prof. Dr. F.W. Asselbergs

f.w.asselbergs@amsterdamumc.nl

020-5664225

Meibergdreef 9

1105 AZ, Amsterdam

Abstract

Background

Several artificial intelligence-enhanced electrocardiogram (AI-ECG) models have shown promise in detecting left ventricular systolic dysfunction (LVSD), but their head-to-head agreement and performance have not been independently compared within the same cohort.

Objectives

To compare the performance of published AI-ECG models for LVSD detection in a standardized external cohort and evaluate the field's transparency and reproducibility.

Methods

We systematically reviewed AI-ECG models predicting LVSD and assessed the risk of bias. Authors were invited to share models for external validation in a well-phenotyped registry of patients undergoing routine clinical cardiac magnetic resonance imaging (CMR) with cardiologist-adjudicated reports and paired ECGs. Model performance was evaluated in all consecutive patients and a lower-complexity subgroup with 15% LVSD prevalence.

Results

We identified 35 studies describing 51 models, reporting high (AUROC >0.80) or excellent (AUROC >0.90) performance. The risk of bias is high and primarily attributed to the limited description of development and validation cohort characteristics, as well as the lack of independent external validation. Four groups (from Korea, the United States, Taiwan, and the Netherlands) shared models for independent testing. AUROCs ranged from 0.83 to 0.93 in all patients (n = 1,203; mean age 59 ± 15 years; 450 [35%] female) and from 0.87 to 0.96 in the lower complexity subset. Performance remained consistent across subgroups, with slight decreases in ECGs showing wide QRS complexes or atrial fibrillation.

Conclusions

In this first-in-kind independent validation and head-to-head comparison study, AI-ECG for LVSD detection demonstrated strong performance despite training on disparate populations. However, the limited availability of models hinders independent validation.

Keywords

Artificial Intelligence, Deep Learning, Electrocardiography, Left Ventricular Dysfunction, Heart Failure

AI-ECG - Artificial intelligence–enhanced electrocardiogram

LVSD - Left ventricular systolic dysfunction

LVEF - Left ventricular ejection fraction

CMR - Cardiac magnetic resonance

TTE - Transthoracic echocardiography

PROBAST - Prediction model Risk Of Bias ASsessment Tool

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses

CGMH - Chang Gung Memorial Hospital

Introduction

Recent advances in artificial intelligence (AI) have enabled the detection of subtle abnormalities in electrocardiograms (ECGs) that often elude expert interpretation.[1,2] This expands the ECG's utility to detect conditions that typically require advanced imaging.[3–7] Given its low cost and widespread availability, AI-enhanced ECG (AI-ECG) models show great promise for large-scale screening and early detection of cardiovascular disease, even in low-resource settings.[8] While the early results of these models are encouraging, clinical adoption remains limited.

A major obstacle to the widespread adoption of AI-ECG is the uncertainty surrounding its performance in real-world settings across diverse populations.[9] To evaluate the generalizability of these models beyond their training population and to identify potential biases introduced by the training population, external validation is crucial.[10,11] However, most AI-ECG models are not publicly available.[10] As a result, external validation is often conducted by the original developers, raising concerns about biases like selective reporting and overfitting to specific populations.[10,12]

Among the various applications of AI-ECG, potentially one of the most clinically impactful and commonly studied is the detection of left ventricular systolic dysfunction (LVSD).[2,13,14] Early and accurate detection of LVSD can significantly improve patient outcomes by enabling timely intervention and management.[15–17] However, LVSD is often diagnosed at an advanced disease stage, which delays the timely initiation of therapy.[18] Using AI-ECG models may provide a straightforward tool for screening patients at risk for LVSD who require further clinical evaluation.[19]

We sought to systematically compare the performance of AI-ECG models for LVSD detection in a standardized external cohort, while assessing the transparency and reproducibility of the field of AI-ECG. We identified all AI-ECG models through a systematic review and evaluated their internal and external performance, as well as the risk of bias of the identified models. We invited all authors to share their models for the first independent external validation and head-to-head comparison of AI-ECG for LVSD detection.

Methods

The institutional review board (Medisch Etische Toetsingscommissie Amsterdam UMC) waived the requirement for informed consent since the study involved secondary analysis of existing data.

Systematic review of AI-ECG for LVSD

To identify all available AI-ECG models for detecting LVSD, we conducted a systematic review of the literature. This systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines **(Supplementary Table 1)**.[20] Screening and data extraction were performed by two independent reviewers (PMC, MJB), and disagreements were resolved through discussion. The comprehensive methods for the systematic review are presented in the Supplementary Materials.

External validation cohort

Data for this study were obtained from a prospectively maintained registry that includes all consecutive patients who underwent cardiac magnetic resonance imaging (CMR) as part of routine clinical care, the Adam CMR database (Amsterdam UMC, The Netherlands). The database contains over 130 variables, including patient demographics, clinical characteristics, and detailed quantitative CMR-image measurements. All CMR scans were conducted between 2015 and 2022, and the reports were reviewed by expert cardiologists. From all patients in the Adam CMR database, clinically obtained 12-lead ECGs (Philips PageWriter or GE MUSE) were collected. All ECGs recorded within a 45-day window, covering 15 days before and 30 days after CMR, were included in the external validation cohort. LVEF was documented as a continuous variable for all patients and used to define the LVEF threshold labels.

To assess the models' behavior in a lower complexity scenario, such as in the cardiology outpatient clinic or emergency department, we created a subcohort, referred to as the representative

HF cohort. This cohort was constructed by stratified resampling from the full CMR cohort to include 15% of patients with LVEF <40% and 10% with LVEF between 40–50. While this cohort is a subset of the CMR cohort and not a real clinical cohort, the prevalence is in accordance with previously published reports of LVSD prevalence in these settings.[21,22] Among the remaining controls, CMR findings were either normal or limited to isolated aortic pathology or coronary artery disease, limiting the number of complex cardiomyopathies with a normal LVEF. This stratification was performed to provide a sensitivity analysis to assess the models' performances in a less complex setting and at the point of care where AI-ECG models are most likely to be applied, such as the emergency department or outpatient clinic. Additionally, since outcome prevalence can significantly impact model performance metrics, this design allows for the assessment of the influence of LVEF distribution on the discriminative performance. The use of CMR, known for its high precision in measuring LVEF, ensures robust ground truth labels, making it a particularly suitable modality for the AI-ECG evaluation.

External validation of included LVSD models

To gain access to existing AI-ECG models, we contacted all the corresponding authors of identified AI-ECG for LVSD publications. Initial emails were sent on July 16, 2024, outlining our request for independent external validation of their models. Specifically, we asked that the developed AI-ECG models for LVSD detection be shared for local deployment at Amsterdam UMC. If no response was received, a follow-up email was sent two weeks later to increase the likelihood of engagement and provide authors with a fair opportunity to share their models for independent validation. The email templates for the initial invitation and follow-up are provided in the Supplementary Methods. For all responding groups, meetings were held to discuss local technical deployment and concerns regarding data and model privacy. After these meetings, the research groups prepared their models for local deployment at Amsterdam UMC.

Model performance assessment

All shared models were applied to the ECGs in the CMR cohort to generate predicted probabilities for LVSD ranging from 0 to 1. The models were delivered as modules; either containerized using Docker or through repositories containing the entire pipeline and model weights. Thresholds were based on those provided by the original developers. Each module ingested raw signals since all signal-processing steps were integrated within the module. The number of ECGs for which models failed to produce valid predictions, for example, due to noise, was recorded and reported for each model. To ensure a fair comparison across models, sensitivity analyses were conducted using only the subset of ECGs that returned valid predictions for each model. Results were stratified by sex (male and female), age (<60 and ≥60 years), QRS duration (<100 ms and ≥100 ms), heart rate (>80 bpm and ≤80 bpm), atrial fibrillation status (present or absent), and clinical diagnosis related to the CMR, including coronary artery disease, ischemia, dilated cardiomyopathy, hypertrophic cardiomyopathy, myocarditis, and other cardiomyopathies. Additionally, false positive rates were calculated for LVEF 40-50% or LVEF >50%.

Sensitivity analyses

To determine if recalibrating the model thresholds could improve model performance in the Amsterdam UMC CMR dataset, we optimized the model probability threshold based on Youden's Index within the Amsterdam UMC dataset. To assess the similarity of the outputs produced by the individual models, we calculated pairwise correlation coefficients using the predicted probabilities of each model.

Statistical analysis

The characteristics of the study population were summarized using means and standard deviations (SD) for continuous variables and counts and percentages for categorical variables. Model performance was evaluated by calculating the area under the receiver operating characteristic curve

(AUROC), sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV), along with 95% confidence intervals (CIs) derived from 500 bootstrap samples. For the representative HF cohort, a random distinct subset of the full cohort was used in each bootstrap iteration, fulfilling the pre-defined distributions. Calibration was evaluated by comparing predicted and observed risks across deciles of predicted probability. Pairwise correlations between model predictions were estimated using Spearman's rank correlation coefficients. Analyses were conducted in Python 3.11.

Results

Identification and review of AI-ECG models

A systematic review of the literature identified 35 eligible studies describing 51 unique AI-ECG models for LVSD **(Table 1)**. The full results of the review are presented in the Supplementary Results, Supplementary Figure S1, and Supplementary Tables S4–S7. Most models were convolutional neural networks trained on ECG signal data, with LVSD mainly defined as LVEF <40% on echocardiography. Among these articles, 9 (25%) reported internal validation results, 11 (31%) included both internal and external validation, and 15 (43%) involved external validation only. All external validation efforts were carried out by the model developers.

Risk of bias of AI-ECG models

The description of baseline characteristics varied, with multiple studies not providing detailed cohort characteristics for both the development and external validation cohorts **(Figure 1A)**. In 12 of the 20 articles describing model derivation, external validation was reported either within the same article or in a separate article (Supplementary Table 6). For the internal validation, the performance was excellent in 29 (57%) of the models, good in 19 (37%), moderate in two (4%), and poor in one (2%), respectively **(Supplementary Table 6)**. For external validation, the performance was excellent in 31 (50%) of the models, good in 28 (45%), and moderate in three (5%), with none being poor **(Figure 1B, Supplementary Table 7)**. Of the articles reporting on the derivation of their models, nine stratified the results to identify potential bias toward subgroups based on age, sex, ethnicity, comorbidity, or ECG characteristics **(Supplementary Table 6)**.

According to the PROBAST assessment, the majority of derivation studies were rated as having high or unclear risk of bias, mainly due to a lack of external validation or inadequate cohort descriptions **(Supplementary Figures 2 and 3)**. Concerns about applicability were infrequent **(Supplementary Figures S2–S3)**.

Model availability

None of the reported models and/or weights were publicly available for independent external validation. After contacting all corresponding authors of articles describing the development of AI-ECG models for LVSD detection, six research groups responded to our inquiries. Of these, four research groups agreed to participate in the external validation study by sharing their models for external validation. The models originated from Korea, the United States (US), Taiwan, and the Netherlands.

The research group identified through Cho et al. shared an algorithm called AiTiALVSD v2.00.00, which is based on a transformer architecture and convolutional encoders.[23] It was trained on 498,726 ECGs from 186,889 patients across 16 tertiary hospitals in Korea, using an 80–10–10%

split. Labels were defined as LVEF <40% on TTE measurement within 30 days of the included ECG. External validation performance of this version of the model has not yet been published.

Sangha et al. introduced an AI-ECG model called ECG Vision.[2] This model utilized a pre-trained version of EfficientNet-B3, which was fine-tuned using a dataset comprising 116,210 patients from Yale New Haven Hospital in the US, with an 85-5-10% split for training, validation, and testing. The model inputs are 300x300 ECG images, either directly or plotted from raw signals. LVSD was defined as LVEF <40% based on TTE measurements performed within 15 days of the ECG. The model achieved an AUROC of 0.90–0.95 across six external validation sites (Supplementary Tables 6 and 7).

Huang et al. shared a ResNet-based model trained on signal data from 380,675 patients at Chang Gung Memorial Hospital (CGMH) in Taiwan, using a 35-15-50% split for training, validation, and testing, which will be referred to as the CGMH model.[13] LVSD was defined as LVEF <40% on TTE performed within two weeks of the ECG. External validation was performed on data from another hospital in Taiwan, resulting in an AUROC of 0.94 for internal validation and 0.95 for external validation **(Supplementary Tables 6 and 7)**.

The research group, as identified through the publication by van de Leur et al.[24] , shared an unpublished ResNet-based model trained on median beat ECGs from 57,250 patients across two hospitals in the Netherlands. They used a 90-10% split for training and internal validation and evaluated it on a different internal cohort of newly referred cardiology outpatients. This model differs from the one in the original publication, with differences in architecture, training, and development cohort. The pipeline applied a custom algorithm to construct median-beat ECGs. As median-beat reconstruction requires consistent and noise-free beats, this preprocessing step was more sensitive to signal noise, leading to rejection of 6% of ECGs. LVSD was defined as LVEF <40% on TTEs performed within 90 days of the ECG.

External validation of AI ECG models for LVSD detection

Independent external validation was conducted using the full CMR cohort, which included 1,203 patients and 4,737 corresponding ECGs, with one randomly selected ECG per individual in each bootstrap iteration. The average age of the cohort was 59 ± 15 years, and 431 (36%) were female. Based on CMR measurements, left ventricular ejection fraction (LVEF) was <40% in 212 (18%) patients, between 40% and 50% in 231 (19%) patients, and >50% in the remaining 760 (63%) patients **(Table 2).** The subset of patients for which all models provided valid predictions included 94% of all ECGs. For the representative heart failure (HF) cohort, 554 patients were included in each bootstrap iteration, consisting of 84 (15%) with LVEF <40%, 50 (9%) with LVEF between 40–50%, and the remaining 420 (76%) with LVEF >50%.

Performance of AI-ECG Models in Detecting Left Ventricular Dysfunction

AiTiALVSD achieved an AUROC of 0.94 (CI: 0.93;0.94) in the full CMR cohort and 0.96 (CI: 0.94;0.97) in the representative HF cohort **(Figure 2A, Supplementary Tables 8-9).** ECG Vision achieved an AUROC of 0.89 (CI: 0.88;0.90) in the full CMR cohort and 0.92 (CI: 0.89;0.94) in the representative HF cohort. The AUROC for CGMH was 0.90 (CI:0.89;0.91) in the full CMR cohort and 0.94 (CI:0.92;0.95) in the representative HF cohort, while the Utrecht model achieved an AUROC of 0.83 (CI: 0.82;0.84) and 0.86 (CI: 0.83;0.89) in these cohorts, respectively. Sensitivity analyses limited to ECGs with valid predictions across all models demonstrated no meaningful change in model performance **(Supplementary Table 10)**.

The AiTiALVSD and ECG Vision slightly underestimated the observed risk, while the models CGMH and Utrecht slightly overestimated it **(Figure 2B)**. Recalibrating the thresholds caused small differences in sensitivity and specificity. In three models, sensitivity improved slightly at the cost of a minor decrease in specificity, whereas in one model, specificity increased slightly with a corresponding drop in sensitivity **(Supplementary Table 11)**.

Stratification by age and sex did not reveal any potential bias **(Figure 3, Supplementary Tables 8 and 9)**. All four models demonstrated slightly reduced performance in ECGs with a QRS width greater than 100ms and in patients with atrial fibrillation. Additionally, AUROCs were generally lower in subgroups with cardiomyopathy. Despite these differences, the models demonstrated good to excellent performance across all subgroups **(Figure 3)**. The false positive rates ranged from 53% to 65% in patients with an LVEF between 40% and 50%, and from 12% to 39% in patients with an LVEF greater than 50% for the included models **(Supplementary Table 12)**.

Correlation Between Individual Models

Pairwise Spearman's correlation coefficients of the continuous model outputs demonstrated moderate to high correlations across all models **(Figure 4)**. The highest correlation was observed between the CGMH model and AiTiALVSD ($\rho = 0.86$). Correlations with the Utrecht model were generally lower (ρ 0.69–0.73).

Discussion

We are the first to introduce a new framework for the external validation of AI-ECG, providing a systematic, head-to-head comparison of all available AI-ECG models for LVSD detection. We found that the overall risk of bias for AI-ECG remains high, mainly due to incomplete descriptions of training cohorts and an overreliance on developer-led validations. Importantly, of the 51 models we identified, only four were accessible for independent testing, highlighting significant gaps in transparency and raising important questions about external validity. The four models available for external validation on a single, well-characterized dataset performed consistently despite differences in training populations from different continents. Our research shows that their diagnostic accuracy depends on edge-case examples, notably individuals with an LVEF between 40 and 50%, where false positives are common. High inter-model correlations indicate that AI-ECG

models are capturing similar electrocardiographic signs of LVSD. Furthermore, validation in a more complex cohort attenuates the diagnostic discrimination ability, underscoring the importance of using representative validation populations.

The strong performance of AI-ECG models for LVSD detection in the literature underscores their potential to screen patients for early diagnosis of LVSD. However, this study reveals significant gaps in the evaluation of AI-ECG models, particularly in assessing potential biases, which may undermine their trustworthiness and hinder clinical adoption. External validation in all reviewed articles was performed solely by the original developers, increasing the risk of selective data inclusion and biased reporting. Moreover, most models do not fully align with the FUTURE-AI guidelines, particularly in terms of transparency, reproducibility, and open access.[10] With the recent FDA approval of AI-ECG models, which signals their potential integration into clinical practice, it is crucial to evaluate the models' performance while minimizing potential bias. This makes it even more concerning that only a small number of AI-ECG models were made available for our external validation study. Selective sharing of models likely introduces systematic bias, as investigators who are more confident in their models may be more willing to share them, creating an inflated and potentially misleading perception of performance at the field level.

The four models shared for independent external validation demonstrated consistent performance, with a slight attenuation of the AUROC when comparing the lower-complexity (representative HF) cohort to the higher-complexity (full CMR) cohort. Notably, this external performance remained consistent, even though three of the models were trained on geographically distinct cohorts (Korea, the US, and Taiwan). Moreover, the continuous probabilities produced by these models exhibited high pairwise correlation, indicating that AI-identified ECG features are consistent despite being trained on three continents, which supports their generalizability. Additionally, models trained on raw ECG signals or ECG images performed similarly well, aligning with recent literature findings.[8,25–27] It is essential to note that all included AI-ECG models were initially developed and validated using TTE-derived LVEF as the ground truth. In contrast, our

independent validation utilized CMR, the reference standard for LVEF assessment, known for its superior accuracy and reproducibility. Despite the differences between these modalities, with TTE tending to underestimate LVEF compared to CMR, the models maintained strong performance, highlighting their robustness and potential clinical utility. The slight performance decline observed in the full CMR cohort warrants careful consideration and possibly model recalibration when implemented in highly complex patient populations.

Our findings demonstrate that the characteristics of the validation cohort substantially influence model performance. An AUROC difference of ~0.04 between the low-complexity (representative HF) and high-complexity (CMR referral) populations underscores this effect. Many published models fail to clearly describe their cohort composition, which limits their interpretability and applicability.[28] Performance variation across subgroups, such as those with broad QRS or atrial fibrillation, further highlights the need for transparent reporting and potential model recalibration.[29] Notably, most false positives in our study had CMR-derived LVEF between 40–50%, suggesting borderline dysfunction and reflecting the challenge of distinguishing subclinical or overlapping cardiac conditions.

Limitations and Future Directions

We included four models for external validation, all of which showed good or excellent performance in our external validation. However, this might have introduced publication bias, as research groups with better-performing models may have been more willing to share them. Additionally, some of the research groups in the external validation provided updated versions of their models that differ from those in the systematic review, precluding direct comparison of their internal and external performance. Furthermore, even though all models were trained to detect LVSD defined as LVEF <40%, heterogeneity in the input format of the ECG and development pipelines limits direct equivalence of performance estimates. Another important limitation is that the analysis was limited to a single Western European cohort of patients referred for CMR,

representing a more challenging classification task and enrichment of complex cardiomyopathies. To mitigate the effect of cohort-specific selection bias, we simulated a population with a prevalence similar to that in the emergency departments or outpatient clinics. This analysis should be interpreted as a sensitivity analysis since it is an augmented cohort, and not a real clinical cohort. Finally, we used CMR-derived LVSD to define the outcome of interest. While CMR is considered the reference standard for LVEF assessment, it systematically differs from TTE-derived LVEF, which was used during model development. While CMR allows for a more accurate evaluation, this modality mismatch may have influenced model performance, particularly in patients with borderline LVEF values.

Future research should prioritize collaborative efforts, sharing models, and independent external validation across datasets that are geographically and demographically diverse. Additionally, using standardized metrics and benchmarks to describe dataset features will be crucial for ensuring consistency and comparability in the literature. With these combined efforts, AI-ECG models can evolve from promising research tools to actual clinical decision-support systems, providing real benefits to patients.

Conclusion

AI-ECG models for detecting LVSD demonstrated high discriminative performance in a novel framework for independent external validation. However, the incomplete reporting of development and validation cohort characteristics in many studies raises concerns about potential bias, and the limited availability of models greatly hinders independent validation, thereby impeding wider clinical adoption.

Author contributions: PMC and MJB designed the study. PMC and MJB performed the study selection and data extraction and drafted the manuscript. PMC performed the analyses and implemented the models. MMW, FWA, CPA, JLS provided critical input on the analysis, as well as

the drafted manuscript. BA, LSD, YH, TM, RK, CK, JK, HL, MSL, RRL, ZL, EKO provided their models for external validation and reviewed the final draft of the manuscript for correctness.

Data availability:

Due to the inclusion of patient-level data, the dataset cannot be publicly shared; however, researchers may contact the corresponding author to request access for academic purposes.

References

1. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol*. 2021;18:465–478.

2. Sangha V, Nargesi AA, Dhingra LS, et al. Detection of left ventricular systolic dysfunction from electrocardiographic images. *Circulation*. 2023;148:765–777.

3. Sangha V, Oikonomou EK, Khera R. Artificial Intelligence Applied to Electrocardiographic Images for Scalable Screening of Transthyretin Amyloid Cardiomyopathy. *Cardiovascular Medicine*. 2024.

4. Khunte A, Sangha V, Oikonomou EK, et al. Automated diagnostic reports from images of electrocardiograms at the Point-of-care. *medRxiv*. 2024. Published onlineFebruary 18, 2024. https://doi.org/10.1101/2024.02.17.24302976.

5. Noseworthy PA, Attia ZI, Behnken EM, et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. *Lancet*. 2022;400:1206–1212.

6. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med*. 2019;25:70–74.

7. Croon PM, Dhingra LS, Biswas D, Oikonomou EK, Khera R. Phenotypic selectivity of artificial intelligence-enhanced electrocardiography in cardiovascular diagnosis and risk prediction. *Circulation*. 2025. Published onlineSeptember 24, 2025. https://doi.org/10.1161/circulationaha.125.076279.

8. Croon PM, Pedroso AF, Khera R. The emerging role of AI in transforming cardiovascular care. *Future Cardiol*. 2025:1–4.

9. Myhre PL, Grenne B, Asch FM, et al. Artificial intelligence-enhanced echocardiography in cardiovascular disease management. *Nat Rev Cardiol*. 2025:1–19.

10. Lekadir K, Frangi AF, Porras AR, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 2025;388:e081554.

11. Yagi R, Goto S, Katsumata Y, MacRae CA, Deo RC. Importance of external validation and subgroup analysis of artificial intelligence in the detection of low ejection fraction from electrocardiograms. *Eur Heart J Digit Health*. 2022;3:654–657.

12. Rosenbaum L. Reconnecting the dots — reinterpreting industry–physician relations. *N Engl J Med*. 2015;372:1860–1864.

13. Huang Y-C, Hsu Y-C, Liu Z-Y, et al. Artificial intelligence-enabled electrocardiographic screening for left ventricular systolic dysfunction and mortality risk prediction. *Front Cardiovasc Med*. 2023;10:1070641.

14. Sandeep B, Huang X, Xiao Z. Artificial intelligence in heart failure improving the efficiency or dependency on it? Letter regarding the article “Artificial intelligence and heart failure: A state-of-the-art review.” *Eur J Heart Fail*. 2024;26:704.

15. Ledwidge M, Gallagher J, Conlon C, et al. Natriuretic peptide-based screening and collaborative care for heart failure: the STOP-HF randomized trial. *JAMA*. 2013;310.

16. Anon. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure Accessed January 21, 2025. https://www.escardio.org/Guidelines/Clinical-Practice-Guidelines/Acute-and-Chronic-Heart-Failure.

17. Heidenreich PA, Bozkurt B, Aguilar D, et al. 2022 AHA/ACC/HFSA guideline for the Management of Heart Failure: A report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. *Circulation*. 2022;145:e895–e1032.

18. Clark KAA, Reinhardt SW, Chouairi F, et al. Trends in heart failure hospitalizations in the US from 2008 to 2018. *J Card Fail*. 2022;28:171–180.

19. Dhingra LS, Aminorroaya A, Sangha V, et al. Heart failure risk stratification using artificial intelligence applied to electrocardiogram images: a multinational study. *Eur Heart J*. 2025;46:1044–1053.

20. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.

21. Gottlieb M, Schraft E, O'Brien J, Patel D, Peksa GD. Prevalence of undiagnosed stage B heart failure among emergency department patients. *Am J Emerg Med*. 2024;85:153–157.

22. Balderston JR, Gertz ZM, Brooks S, Joyce JM, Evans DP. Diagnostic yield and accuracy of bedside echocardiography in the emergency department in hemodynamically stable patients. *J Ultrasound Med*. 2019;38:2845–2851.

23. Cho J, Lee B, Kwon J-M, et al. Artificial intelligence algorithm for screening heart failure with reduced ejection fraction using electrocardiography. *ASAIO J*. 2021;67:314–321.

24. van de Leur RR, Bos MN, Taha K, et al. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *Eur Heart J Digit Health*. 2022;3:390–404.

25. Sau A, Zeidaabadi B, Patlatzoglou K, et al. A comparison of artificial intelligence-enhanced electrocardiography approaches for prediction of time-to-mortality using electrocardiogram images. *Eur Heart J Digit Health*. 2024:ztae090.

26. Sangha V, Dhingra LS, Aminorroaya A, et al. Identification of hypertrophic cardiomyopathy on electrocardiographic images with deep learning. *Nat Cardiovasc Res*. 2025:1–10.

27. Dhingra LS, Aminorroaya A, Sangha V, et al. Ensemble deep learning algorithm for structural heart disease screening using electrocardiographic images. *J Am Coll Cardiol*. 2025;85:1302–1313.

28. Galanty M, Luitse D, Noteboom SH, et al. Assessing the documentation of publicly available medical image and signal datasets and their impact on bias using the BEAMRAD tool. *Sci Rep*. 2024;14:31846.

29. Wu J, Biswas D, Ryan M, et al. Artificial intelligence methods for improved detection of undiagnosed heart failure with preserved ejection fraction. *Eur J Heart Fail*. 2024;26:302–310.

30. Bergquist JA, Zenger B, Brundage J, et al. Performance of off-the-shelf machine learning architectures and biases in low left ventricular ejection fraction detection. *Heart Rhythm O2*. 2024;5:644–654.

31. Khunte A, Sangha V, Oikonomou EK, et al. Detection of left ventricular systolic dysfunction from single-lead electrocardiography adapted for portable and wearable devices. *NPJ Digit Med*. 2023;6:124.

32. Vaid A, Jiang J, Sawant A, et al. A foundational vision transformer improves diagnostic performance for electrocardiograms. *NPJ Digit Med*. 2023;6:108.

33. Honarvar H, Agarwal C, Somani S, et al. Enhancing convolutional neural network predictions of electrocardiograms with left ventricular dysfunction using a novel sub-waveform representation. *Cardiovasc Digit Health J*. 2022;3:220–231.

34. Chen H-Y, Lin C-S, Fang W-H, 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 (Lausanne)*. 2022;9:870523.

35. Golany T, Radinsky K, Kofman N, et al. Physicians and machine-learning algorithm performance in predicting left-ventricular systolic dysfunction from a standard 12-lead-electrocardiogram. *J Clin Med*. 2022;11:6767.

36. Lee C-H, Liu W-T, Lou Y-S, et al. Artificial intelligence-enabled electrocardiogram screens low left ventricular ejection fraction with a degree of confidence. *Digit Health*. 2022;8:20552076221143250.

37. Chen H-Y, Lin C-S, Fang W-H, et al. Artificial intelligence-enabled electrocardiography predicts left ventricular dysfunction and future cardiovascular outcomes: A retrospective analysis. *J Pers Med*. 2022;12:455.

38. Kwon J-M, Jo Y-Y, Lee SY, et al. Artificial intelligence-enhanced smartwatch ECG for heart failure-reduced ejection fraction detection by generating 12-lead ECG. *Diagnostics (Basel)*. 2022;12:654.

39. Surendra K, Nürnberg S, Bremer JP, et al. Pragmatic screening for heart failure in the general population using an electrocardiogram-based neural network. *ESC Heart Fail*. 2023;10:975–984.

40. Vaid A, Johnson KW, Badgeley MA, et al. Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. *JACC Cardiovasc Imaging*. 2022;15:395–410.

41. Katsushika S, Kodera S, Nakamoto M, et al. The effectiveness of a deep learning model to detect left ventricular systolic dysfunction from electrocardiograms. *Int Heart J*. 2021;62:1332–1341.

42. Choi J, Lee S, Chang M, Lee Y, Oh GC, Lee H-Y. Deep learning of ECG waveforms for diagnosis of heart failure with a reduced left ventricular ejection fraction. *Sci Rep*. 2022;12:14235.

43. Sun J-Y, Qiu Y, Guo H-C, et al. A method to screen left ventricular dysfunction through ECG based on convolutional neural network. *J Cardiovasc Electrophysiol*. 2021;32:1095–1102.

44. Attia ZI, Harmon DM, Dugan J, et al. Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction. *Nat Med*. 2022;28:2497–2503.

45. Harmon DM, Adedinsewo D, Van't Hof JR, et al. Community-based participatory research application of an artificial intelligence-enhanced electrocardiogram for cardiovascular disease screening: A FAITH! Trial ancillary study. *Am J Prev Cardiol*. 2022;12:100431.

46. Harmon DM, Carter RE, Cohen-Shelly M, et al. Real-world performance, long-term efficacy, and absence of bias in the artificial intelligence enhanced electrocardiogram to detect left ventricular systolic dysfunction. *Eur Heart J Digit Health*. 2022;3:238–244.

47. Attia ZI, Dugan J, Rideout A, et al. Automated detection of low ejection fraction from a one-lead electrocardiogram: application of an AI algorithm to an electrocardiogram-enabled Digital Stethoscope. *Eur Heart J Digit Health*. 2022;3:373–379.

48. Attia IZ, Tseng AS, Benavente ED, et al. External validation of a deep learning electrocardiogram algorithm to detect ventricular dysfunction. *Int J Cardiol*. 2021;329:130–135.

49. Attia ZI, Kapa S, Yao X, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction: ATTIA et al. *J Cardiovasc Electrophysiol*. 2019;30:668–674.

50. Klein CJ, Ozcan I, Attia ZI, et al. Electrocardiogram-artificial intelligence and immune-mediated necrotizing myopathy: Predicting left ventricular dysfunction and clinical outcomes. *Mayo Clin Proc Innov Qual Outcomes*. 2022;6:450–457.

51. Bachtiger P, Petri CF, Scott FE, et al. Point-of-care screening for heart failure with reduced ejection fraction using artificial intelligence during ECG-enabled stethoscope examination in London, UK: a prospective, observational, multicentre study. *Lancet Digit Health*. 2022;4:e117–e125.

52. Kashou AH, Medina-Inojosa JR, Noseworthy PA, et al. Artificial intelligence-augmented electrocardiogram detection of left ventricular systolic dysfunction in the general population. *Mayo Clin Proc*. 2021;96:2576–2586.

53. Brito BO de F, Attia ZI, Martins LNA, et al. Left ventricular systolic dysfunction predicted by artificial intelligence using the electrocardiogram in Chagas disease patients-The SaMi-Trop cohort. *PLoS Negl Trop Dis*. 2021;15:e0009974.

54. Kashou AH, Noseworthy PA, Lopez-Jimenez F, et al. The effect of cardiac rhythm on artificial intelligence-enabled ECG evaluation of left ventricular ejection fraction prediction in cardiac intensive care unit patients. *Int J Cardiol*. 2021;339:54–55.

55. Jentzer JC, Kashou AH, Attia ZI, et al. Left ventricular systolic dysfunction identification using artificial intelligence-augmented electrocardiogram in cardiac intensive care unit patients. *Int J Cardiol*. 2021;326:114–123.

56. Adedinsewo D, Carter RE, Attia Z, et al. Artificial Intelligence-Enabled ECG Algorithm to Identify Patients With Left Ventricular Systolic Dysfunction Presenting to the Emergency Department With Dyspnea. *Circulation: Arrhythmia and Electrophysiology*. 2020. Published onlineAugust 2020. https://doi.org/10.1161/circep.120.008437.

57. Noseworthy PA, Attia ZI, Brewer LC, et al. Assessing and mitigating bias in medical artificial intelligence: The effects of race and ethnicity on a deep learning model for ECG analysis: The effects of race and ethnicity on a deep learning model for ECG analysis. *Circ Arrhythm Electrophysiol*. 2020;13:e007988.

58. Vaid A, Jiang JJ, Sawant A, et al. Automated determination of left ventricular function using electrocardiogram data in patients on maintenance hemodialysis. *Clin J Am Soc Nephrol*. 2022;17:1017–1025.

TABLES AND FIGURES

Table 1: Overview of all included articles

First Author	Year of publication	Outcome label	Validation
Bergquist et al.[30]	2023	LVEF <40%	Internal validation
Khunte et al.[31]	2023	LVEF < 40%	Internal validation
Vaid et al.[32]	2023	LVEF < 40%	Internal and external validation
Honarvar et al.[33]	2022	LVEF <35%	Internal validation
Huang et al.[13]	2023	LVEF < 40%	Internal and external validation
Chen et al.[34]	2022	LVEF <35%, <50%	Internal and external validation
Golany et al.[35]	2022	LVEF <35%	Internal validation
Yagi et al.[11]	2022	LVEF <40%	Internal and external validation
Lee et al.[36]	2022	LVEF <40%	Internal and external validation
Sangha et al.[2]	2023	LVEF <40%	Internal and external validation
Chen et al.[37]	2022	LVEF <35%,	Internal and external validation
Kwon et al.[38]	2022	LVEF <40%	Internal and external validation
Surendra et al.[39]	2023	Asymptomatic LVSD, HFrEF, HFmrEF, HFpEF	Internal validation
Vaid et al.[40]	2022	LVEF <35%	Internal and external validation
Katsushika et al.[41]	2021	LVEF <40%	Internal validation
Cho et al.[23]	2021	LVEF <40%	Internal and external validation
Choi et al.[42]	2022	LVEF <40%	External validation model Cho (2021)
Sun et al.[43]	2021	LVEF <50%	internal validation
Attia et al.[6]	2019	LVEF <35%	internal validation

Attia et al.[44]	2022	LVEF < 40%	External validation model Attia (2019)
Harmon et al.[45]	2022	LVEF < 50%, <40%	External validation model Attia (2019)
Harmon et al.[46]	2022	LVEF <40%	External validation model Attia (2019)
Attia et al.[47]	2022	LVEF < 35%, <40%, <50%	External validation model Attia (2019)
Attia et al.[48]	2021	LVEF <35%	External validation model Attia (2019)
Attia et al.[49]	2019	LVEF < 35%	External validation model Attia (2019)
Klein et al.[50]	2022	LVEF <35%	External validation model Attia (2019)
Bachtiger et al.[51]	2021	LVEF < 40%	External validation model Attia (2019)
Kashou et al.[52]	2021	LVEF <35%	External validation model Attia (2019)
Brito et al.[53]	2021	LVEF <40%	External validation model Attia (2019)
Kashou et al.[54]	2021	LVEF 35%	External validation model Attia (2019)
Jentzer et al.[55]	2021	LVEF <35%	External validation model Attia (2019)
Adedinsewo et al.[56]	2020	LVEF <50%	External validation model Attia (2019)
Noseworthy et al.[57]	2020	LVEF <35%	External validation model Attia (2019)
van de Leur et al.[24]	2022	LVEF <40%	Internal and external validation
Vaid et al.[58]	2022	LVEF <40%	Internal validation

Abbreviations: LVEF, left ventricular ejection fraction; LVSD, left ventricular systolic dysfunction, HFrEF, heart failure with reduced ejection fraction; HFpEF, heart failure with preserved ejection fraction.

Table 2: Baseline characteristics of the Amsterdam UMC CMR-cohort.

Characteristic	**n (%) or mean ± SD**
Number of ECGs	4737
Number of unique Patients	1203
Age	59.5 ± 14.9
Female	431 (35.9)
LVEF 40-50	231 (19.3)
LVEF<40	212 (17.7)
Coronary artery disease	328 (27.6)
Cardiomyopathy other	147 (12.4)
Ischemia	87 (7.3)
Dilated cardiomyopathy	83 (7.0)
Aortic pathology	49 (4.1)
Hypertrophic cardiomyopathy	63 (5.3)

Diagnoses are based on the conclusions of the cardiac magnetic resonance imaging study reported by an expert cardiologist. Abbreviations: n, number of samples; SD, Standard Deviation; LVEF, Left Ventricular Ejection Fraction.

Figure 1: Reported baseline characteristics, internal and external performance for included models.

Panel A: These heatmaps display baseline characteristics reported in derivation (top) and external validation (bottom) studies. Each row represents a study, and each column indicates a specific characteristic (e.g., age, sex, comorbidities). Green shows that the characteristic was reported, while red indicates it was not. Panel B: The difference between internal validation AUROC and external validation AUROC across models. The purple diamonds represent external validation reported in the original paper. The red diamonds show external validation conducted in separate papers. All values are referenced to internal AUROC; therefore, a value higher than zero indicates a higher internal AUROC compared to the AUROC during external validation. Abbreviations: AUROC, Area Under the Receiver Operating Curve; CMR, Cardiac Magnetic Resonance; HF, Heart failure.

Figure 2: Performance and calibration of included models for external validation

Panel A: Forest plot showing the independent external validation results of four models (Cho et al., Sangha et al., Huang et al., and van de Leur et al.) in the full CMR cohort and the augmented heart failure cohort. The plot displays AUROC values with their 95% confidence intervals (CI). Panel B: Calibration plot for the external validation. Mean predicted probabilities versus observed proportions, shown using quantile binning (10 bins). The "Perfect Calibration" line indicates ideal model calibration. Abbreviations: CMR, Cardiac Magnetic Resonance imaging; HF, Heart Failure; AUROC, Area Under the Receiver Operating Characteristic Curve; Sens, Sensitivity; Spec, Specificity; PPV, Positive Predictive Value; NPV, Negative Predictive Value.

Figure 3: Subgroup analysis of AUROC values for external validation of models.

Performance in the full CMR cohort across various patient subgroups is shown, including sex, age, QRS duration, heart rate, atrial fibrillation status, coronary disease, types of cardiomyopathies, and myocarditis. The total number of patients is listed, along with the number of cases with LVSD in the chosen category (n patients (cases EF<40%)). CMR; Cardiac Magnetic Resonance imaging,

AUROC values are provided with 95% confidence intervals to give insights into model performance within specific populations. Abbreviations: AUROC, Area Under the Receiver Operating Characteristic Curve; HCM, Hypertrophic Cardiomyopathy; DCM, Dilated Cardiomyopathy; CMP, Cardiomyopathy.

Figure 4: Correlation between the different AI-ECG models

The heatmap shows the pairwise Spearman correlation coefficients between the model predictions of the different included models. Abbreviations: CMR, Cardiac Magnetic Resonance imaging; HF, Heart Failure; AUROC, Area Under the Receiver Operating Characteristic Curve; Sens, Sensitivity; Spec, Specificity; PPV, Positive Predictive Value; NPV, Negative Predictive Value.

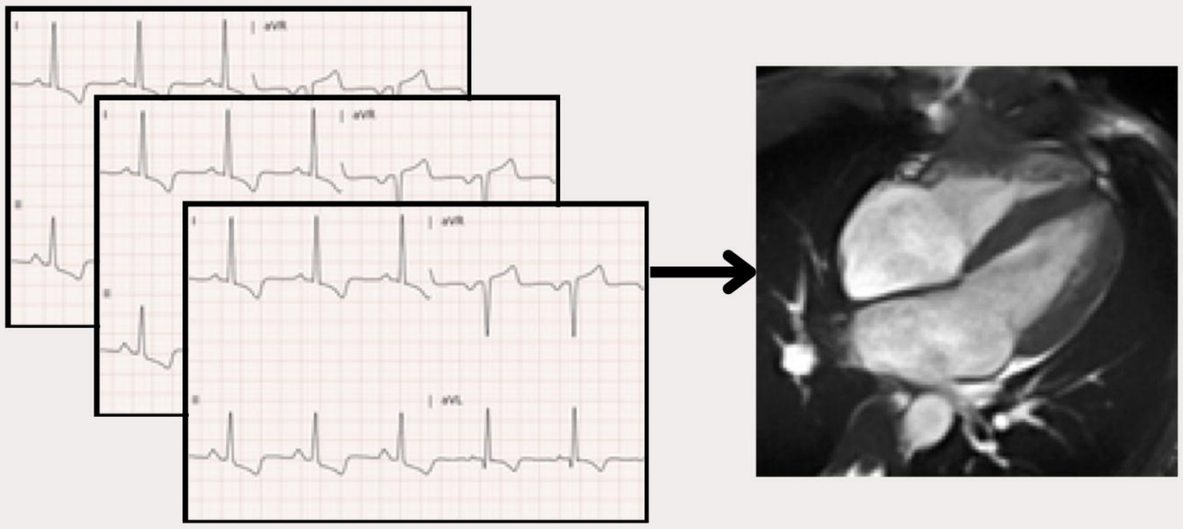
Central Illustration

We propose a structured framework for the independent validation of artificial intelligence–enabled electrocardiography (AI-ECG) models. Of 51 published AI-ECG models for left ventricular systolic dysfunction (LVSD), only four were available for true external validation. These models were evaluated in an independent cohort with paired electrocardiograms and cardiac magnetic resonance imaging (CMR) as the reference standard. Despite differences in architecture and training cohorts, model performance remained high in this external setting. However, limited openness and accessibility of AI-ECG models constrain independent assessment and may impede broader clinical adoption.

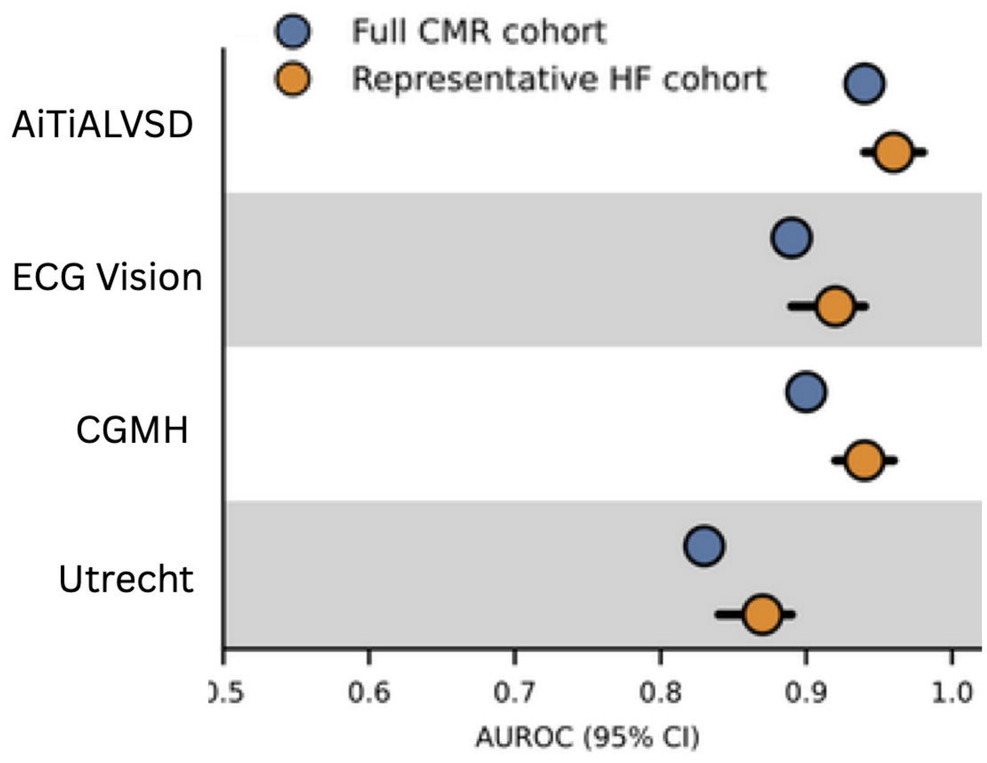
We propose a novel framework for the independent validation of AI-ECG

Out of 51 AI-ECG for LVSD models identified, four were available for external validation

Validation in a cohort with ECGs and paired cardiac magnetic resonance imaging



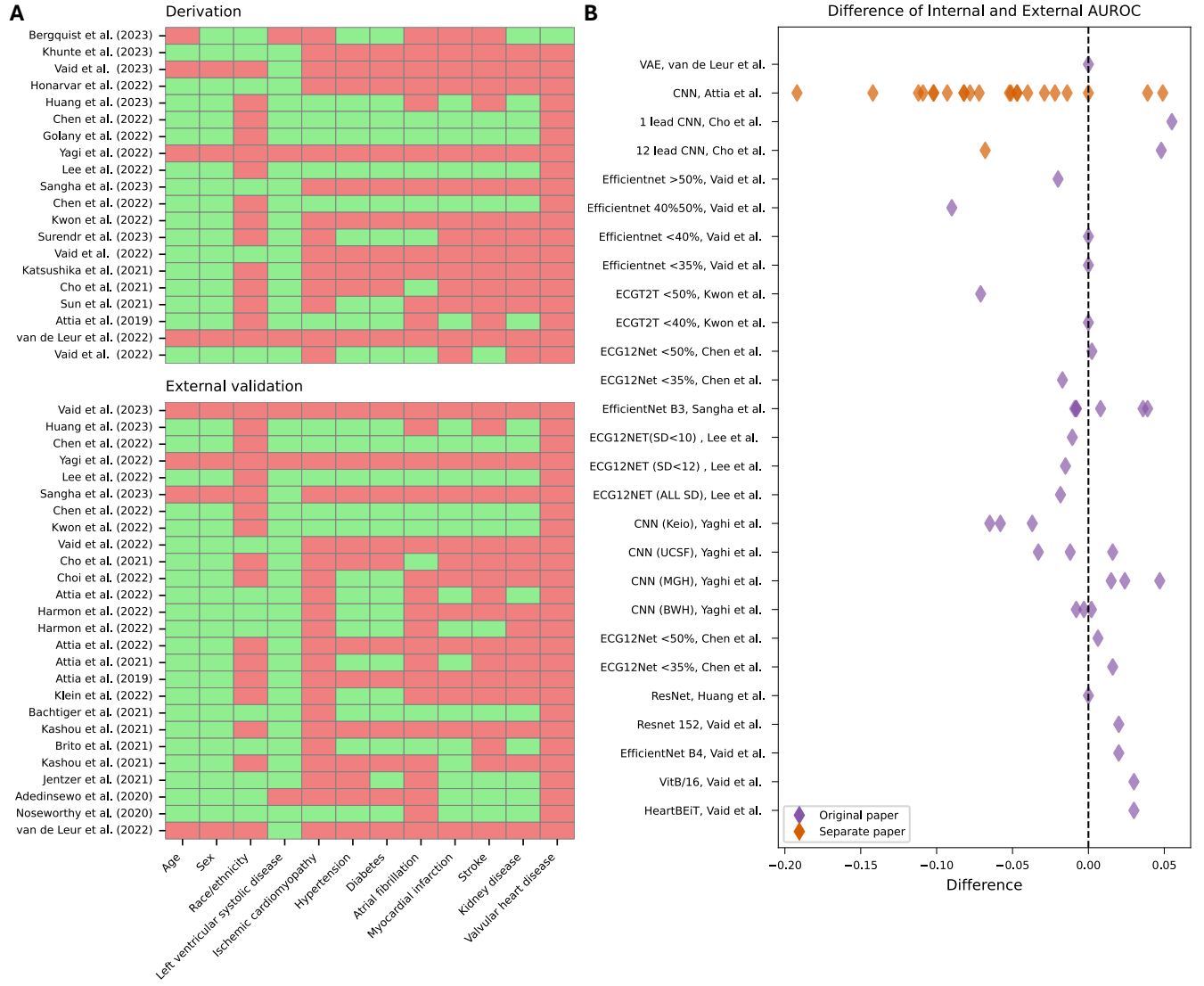
Performance remained excellent

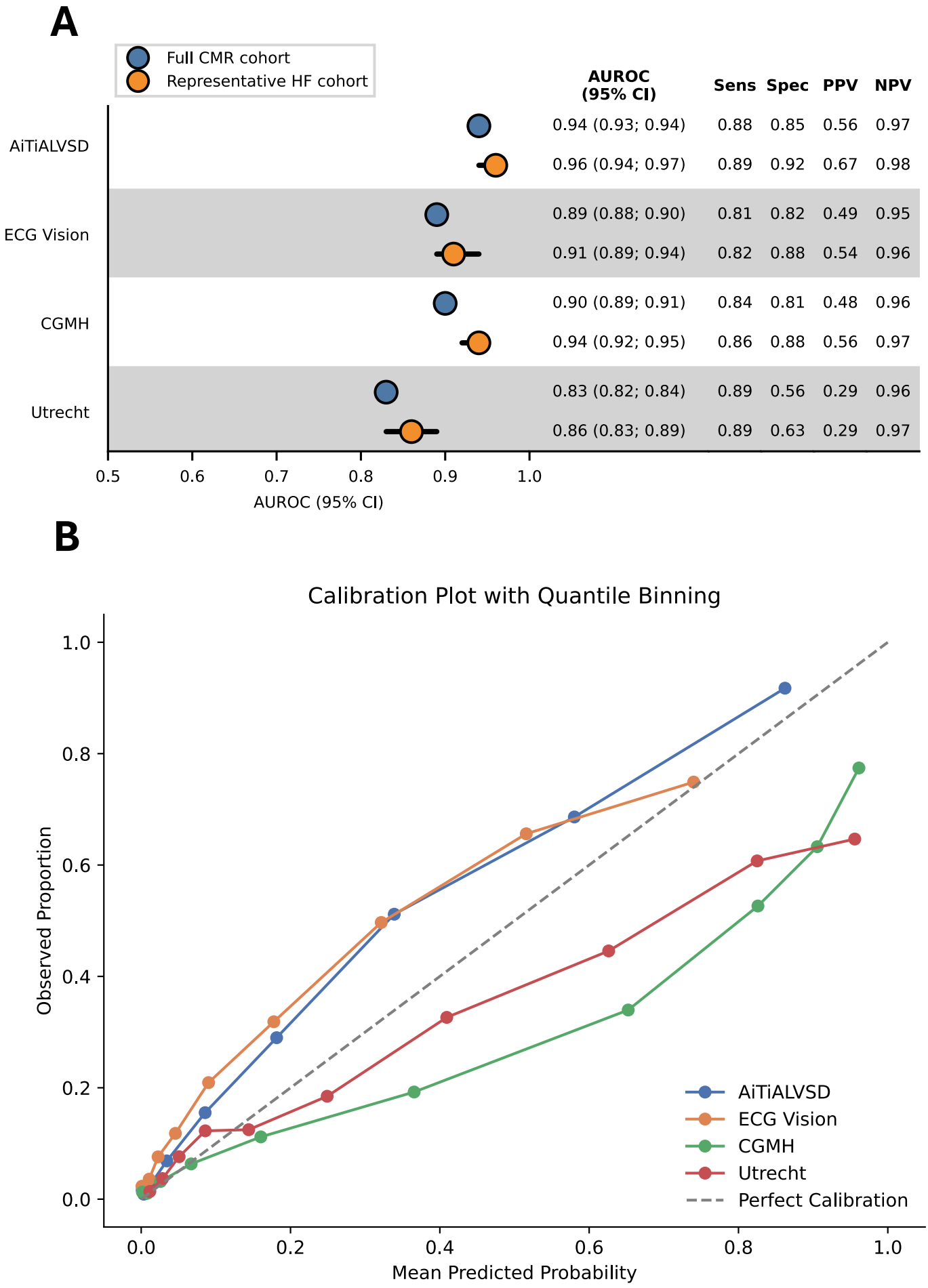


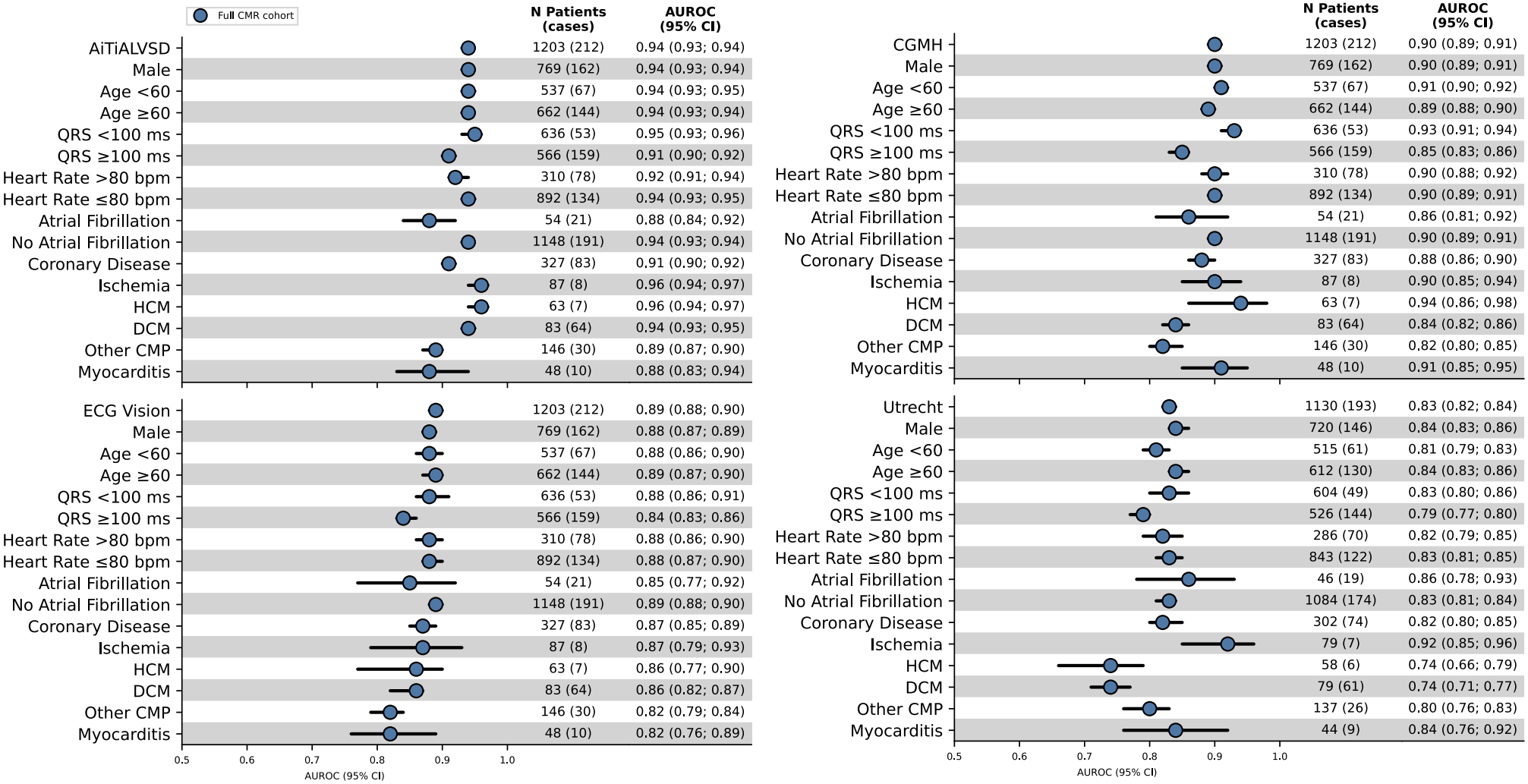
Conclusions

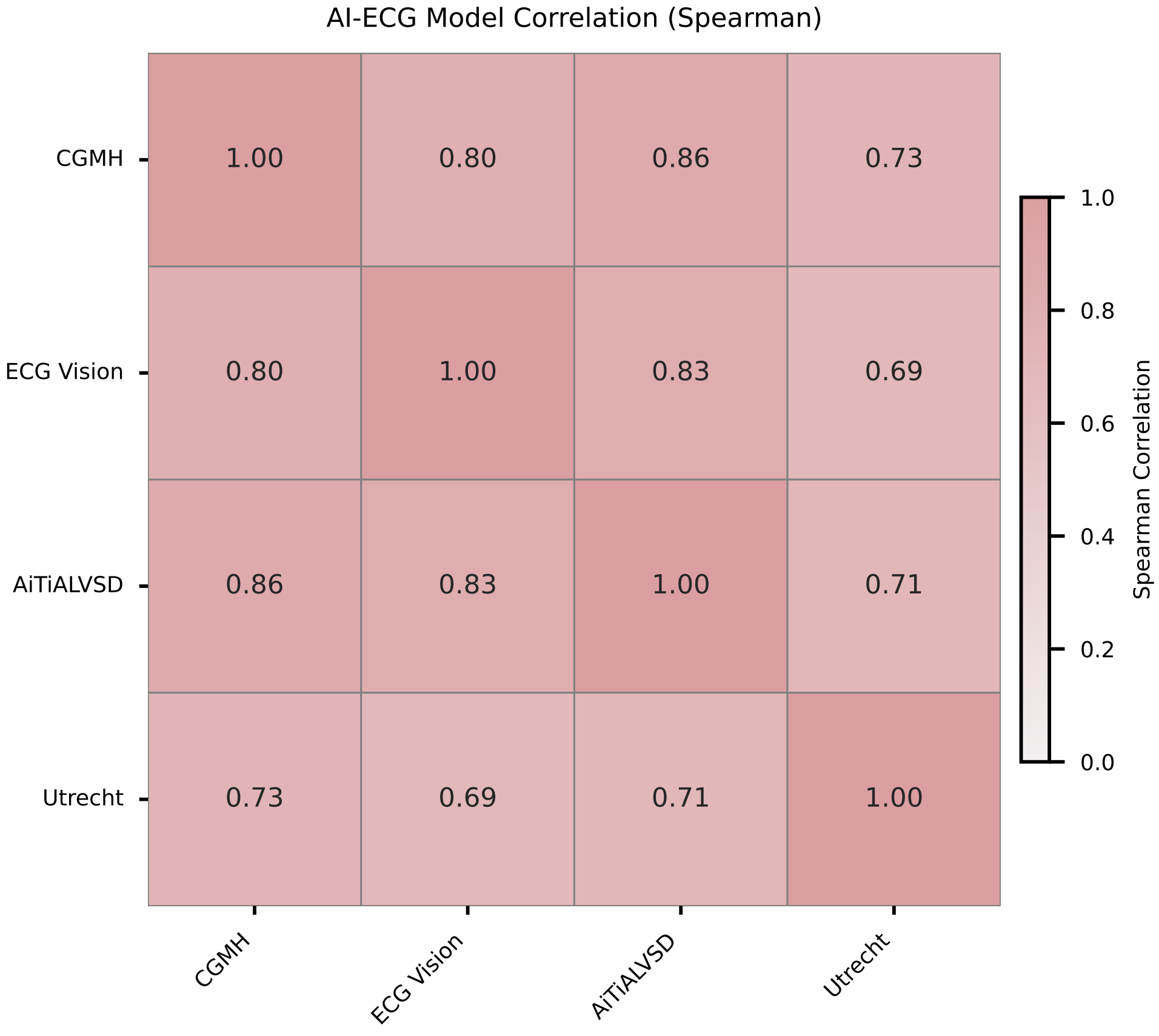
AI-ECG shows strong discrimination in independent external validation using CMR

Limited openness of models limits the independent assessment AI-ECG and impedes further adoption


A
Derivation
Bergquist et al. (2023)
Khunte et al. (2023)
Vaid et al. (2023)
Honarvar et al. (2022)
Huang et al. (2023)
Chen et al. (2022)
Golany et al. (2022)
Yagi et al. (2022)
Lee et al. (2022)
Sangha et al. (2023)
Chen et al. (2022)
Kwon et al. (2022)
Surendr et al. (2023)
Vaid et al. (2022)
Katsushika et al. (2021)
Cho et al. (2021)
Sun et al. (2021)
Attia et al. (2019)
van de Leur et al. (2022)
Vaid et al. (2022)
External validation
Vaid et al. (2023)
Huang et al. (2023)
Chen et al. (2022)
Yagi et al. (2022)
Lee et al. (2022)
Sangha et al. (2023)
Chen et al. (2022)
Kwon et al. (2022)
Vaid et al. (2022)
Cho et al. (2021)
Choi et al. (2022)
Attia et al. (2022)
Harmon et al. (2022)
Harmon et al. (2022)
Attia et al. (2022)
Attia et al. (2021)
Attia et al. (2019)
Klein et al. (2022)
Bachtiger et al. (2021)
Kashou et al. (2021)
Brito et al. (2021)
Kashou et al. (2021)
Jentzer et al. (2021)
Adedinsewo et al. (2020)
Noseworthy et al. (2020)
van de Leur et al. (2022)
Age
Sex
Race/ethnicity
Left ventricular systolic disease
Ischemic cardiomyopathy
Hypertension
Diabetes
Atrial fibrillation
Myocardial infarction
Stroke
Kidney disease
Valvular heart disease
B
Difference of Internal and External AUROC
VAE, van de Leur et al.
CNN, Attia et al.
1 lead CNN, Cho et al.
12 lead CNN, Cho et al.
Efficientnet >50%, Vaid et al.
Efficientnet 40%50%, Vaid et al.
Efficientnet <40%, Vaid et al.
Efficientnet <35%, Vaid et al.
ECGT2T <50%, Kwon et al.
ECGT2T <40%, Kwon et al.
ECG12Net <50%, Chen et al.
ECG12Net <35%, Chen et al.
EfficientNet B3, Sangha et al.
ECG12NET(SD<10) , Lee et al.
ECG12NET (SD<12) , Lee et al.
ECG12NET (ALL SD), Lee et al.
CNN (Keio), Yaghi et al.
CNN (UCSF), Yaghi et al.
CNN (MGH), Yaghi et al.
CNN (BWH), Yaghi et al.
ECG12Net <50%, Chen et al.
ECG12Net <35%, Chen et al.
ResNet, Huang et al.
Resnet 152, Vaid et al.
EfficientNet B4, Vaid et al.
VitB/16, Vaid et al.
HeartBEiT, Vaid et al.
Original paper
Separate paper
−0.20
−0.15
−0.10
−0.05
0.00
0.05
Difference


A
Full CMR cohort
Representative HF cohort
AUROC (95% CI)
Sens
Spec
PPV
NPV
AiTiALVSD
0.94 (0.93; 0.94) 0.88 0.85 0.56 0.97
0.96 (0.94; 0.97) 0.89 0.92 0.67 0.98
ECG Vision
0.89 (0.88; 0.90) 0.81 0.82 0.49 0.95
0.91 (0.89; 0.94) 0.82 0.88 0.54 0.96
CGMH
0.90 (0.89; 0.91) 0.84 0.81 0.48 0.96
0.94 (0.92; 0.95) 0.86 0.88 0.56 0.97
Utrecht
0.83 (0.82; 0.84) 0.89 0.56 0.29 0.96
0.86 (0.83; 0.89) 0.89 0.63 0.29 0.97
0.5
0.6
0.7
0.8
0.9
1.0
AUROC (95% CI)
B
Calibration Plot with Quantile Binning
Observed Proportion
Mean Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0
AiTiALVSD
ECG Vision
CGMH
Utrecht
Perfect Calibration


Full CMR cohort
N Patients (cases)
AUROC (95% CI)
AiTiALVSD 1203 (212) 0.94 (0.93; 0.94)
Male 769 (162) 0.94 (0.93; 0.94)
Age <60 537 (67) 0.94 (0.93; 0.95)
Age ≥60 662 (144) 0.94 (0.93; 0.94)
QRS <100 ms 636 (53) 0.95 (0.93; 0.96)
QRS ≥100 ms 566 (159) 0.91 (0.90; 0.92)
Heart Rate >80 bpm 310 (78) 0.92 (0.91; 0.94)
Heart Rate ≤80 bpm 892 (134) 0.94 (0.93; 0.95)
Atrial Fibrillation 54 (21) 0.88 (0.84; 0.92)
No Atrial Fibrillation 1148 (191) 0.94 (0.93; 0.94)
Coronary Disease 327 (83) 0.91 (0.90; 0.92)
Ischemia 87 (8) 0.96 (0.94; 0.97)
HCM 63 (7) 0.96 (0.94; 0.97)
DCM 83 (64) 0.94 (0.93; 0.95)
Other CMP 146 (30) 0.89 (0.87; 0.90)
Myocarditis 48 (10) 0.88 (0.83; 0.94)
ECG Vision 1203 (212) 0.89 (0.88; 0.90)
Male 769 (162) 0.88 (0.87; 0.89)
Age <60 537 (67) 0.88 (0.86; 0.90)
Age ≥60 662 (144) 0.89 (0.87; 0.90)
QRS <100 ms 636 (53) 0.88 (0.86; 0.91)
QRS ≥100 ms 566 (159) 0.84 (0.83; 0.86)
Heart Rate >80 bpm 310 (78) 0.88 (0.86; 0.90)
Heart Rate ≤80 bpm 892 (134) 0.88 (0.87; 0.90)
Atrial Fibrillation 54 (21) 0.85 (0.77; 0.92)
No Atrial Fibrillation 1148 (191) 0.89 (0.88; 0.90)
Coronary Disease 327 (83) 0.87 (0.85; 0.89)
Ischemia 87 (8) 0.87 (0.79; 0.93)
HCM 63 (7) 0.86 (0.77; 0.90)
DCM 83 (64) 0.86 (0.82; 0.87)
Other CMP 146 (30) 0.82 (0.79; 0.84)
Myocarditis 48 (10) 0.82 (0.76; 0.89)
0.5 0.6 0.7 0.8 0.9 1.0
AUROC (95% CI)
N Patients (cases)
AUROC (95% CI)
CGMH 1203 (212) 0.90 (0.89; 0.91)
Male 769 (162) 0.90 (0.89; 0.91)
Age <60 537 (67) 0.91 (0.90; 0.92)
Age ≥60 662 (144) 0.89 (0.88; 0.90)
QRS <100 ms 636 (53) 0.93 (0.91; 0.94)
QRS ≥100 ms 566 (159) 0.85 (0.83; 0.86)
Heart Rate >80 bpm 310 (78) 0.90 (0.88; 0.92)
Heart Rate ≤80 bpm 892 (134) 0.90 (0.89; 0.91)
Atrial Fibrillation 54 (21) 0.86 (0.81; 0.92)
No Atrial Fibrillation 1148 (191) 0.90 (0.89; 0.91)
Coronary Disease 327 (83) 0.88 (0.86; 0.90)
Ischemia 87 (8) 0.90 (0.85; 0.94)
HCM 63 (7) 0.94 (0.86; 0.98)
DCM 83 (64) 0.84 (0.82; 0.86)
Other CMP 146 (30) 0.82 (0.80; 0.85)
Myocarditis 48 (10) 0.91 (0.85; 0.95)
Utrecht 1130 (193) 0.83 (0.82; 0.84)
Male 720 (146) 0.84 (0.83; 0.86)
Age <60 515 (61) 0.81 (0.79; 0.83)
Age ≥60 612 (130) 0.84 (0.83; 0.86)
QRS <100 ms 604 (49) 0.83 (0.80; 0.86)
QRS ≥100 ms 526 (144) 0.79 (0.77; 0.80)
Heart Rate >80 bpm 286 (70) 0.82 (0.79; 0.85)
Heart Rate ≤80 bpm 843 (122) 0.83 (0.81; 0.85)
Atrial Fibrillation 46 (19) 0.86 (0.78; 0.93)
No Atrial Fibrillation 1084 (174) 0.83 (0.81; 0.84)
Coronary Disease 302 (74) 0.82 (0.80; 0.85)
Ischemia 79 (7) 0.92 (0.85; 0.96)
HCM 58 (6) 0.74 (0.66; 0.79)
DCM 79 (61) 0.74 (0.71; 0.77)
Other CMP 137 (26) 0.80 (0.76; 0.83)
Myocarditis 44 (9) 0.84 (0.76; 0.92)
0.5 0.6 0.7 0.8 0.9 1.0
AUROC (95% CI)


AI-ECG Model Correlation (Spearman)
CGMH
ECG Vision
AiTiALVSD
Utrecht
1.00
0.80
0.86
0.73
0.80
1.00
0.83
0.69
0.86
0.83
1.00
0.71
0.73
0.69
0.71
1.00
CGMH
ECG Vision
AiTiALVSD
Utrecht
Spearman Correlation
1.0
0.8
0.6
0.4
0.2
0.0