AI-ECG for LVSD detection: a systematic review and first-in-kind multinational head-to-head comparison

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

Word count text: 4165

Word count abstract: 249/250

Number of references: 46

Number of tables and figures: 6

Correspondence:

Prof. Dr. F.W. Asselbergs

f.w.asselbergs@amsterdamumc.nl

Meibergdreef 9

1105 AZ, Amsterdam

Abstract

Background and aims

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 evaluated within the same cohort. This study aimed to systematically characterize available AI-ECG models for LVSD detection, and for the first time evaluate their head-to head performance.

Methods

We systematically reviewed published AI-ECG models for LVSD detection, extracting data on architecture, cohort characteristics, and performance. Authors of these studies were invited to share models for external validation in a deeply phenotyped registry of all patients undergoing cardiac magnetic resonance imaging (CMR) in the Amsterdam UMC with paired ECGs. Performance was assessed in all consecutive patients and in a subset reflecting a lower-complexity sample (15% prevalence of LVEF <40%).

Results

We identified 35 studies describing 51 AI-ECG models, all showing high performance (area under the receiver operating characteristic curve [AUROC]>80) or excellent (AUROC >90) performance. Four research groups (Korea, the United States, Taiwan, The Netherlands) shared models for external validation. AUROCs ranged from 0.83–0.93 in all consecutive patients (n = 1,306; mean age 59 ± 15 years; 450 [35%] female) and 0.87–0.96 in the lower complexity subset. Performance remained consistent across subgroups, with minor degradations in ECGs displaying a wide QRS-complex or atrial fibrillation.

Conclusions

In this first-in-kind independent validation and head-to-head comparison study, AI-ECG for LVSD detection exhibited robust performance despite disparate training populations. However, limited availability of models impedes independent validation and clinical adoption.

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

Introduction

The electrocardiogram (ECG) is the cornerstone for the detection of several cardiac diagnoses by non-invasively recording the electrical activity of the heart. Recent advances in artificial intelligence (AI) have enabled the detection of subtle abnormalities that often elude expert interpretation, expanding the ECG's utility to conditions typically requiring advanced imaging.[1–4] Given its low cost and widespread availability, AI-enhanced ECG (AI-ECG) models hold great promise for scalable screening and early detection of cardiovascular disease, including in low-resource settings.[5]

The rise of big data and advances in computing have enabled the development of numerous AI-ECG models for cardiovascular care. While early results are promising, clinical adoption remains limited. A key barrier is uncertainty about real-world performance across diverse populations. To address these concerns, the FUTURE-AI guidelines[6] were established to provide a comprehensive framework for the responsible development and evaluation of AI in healthcare, emphasizing key principles such as transparency, fairness, and robustness to promote trustworthy and clinically applicable AI systems. One of the most important items to ensure these key principles, both in these guidelines and other literature, is external validation. External validation is used to assess generalizability of performance and identify potential biases introduced by the training population.[6,7] However, most AI-ECG models are not made publicly available, and consequently external validation is often conducted by the original developers, with concerns about biases such as selective reporting and overfitting to specific populations.[5]

Among the various applications of AI-ECG, one of the most clinically impactful and commonly studied is the detection of left ventricular systolic dysfunction (LVSD).[3,8] Early and accurate detection of LVSD can significantly improve patient outcomes by enabling timely intervention and management.[9–11] However, LVSD is often diagnosed in an advanced disease stage, therewith delaying timely initiation of therapy. Using AI-ECG models may offer an easy to apply tool to screen patients at risk for LVSD requiring further clinical workup.

This systematic review provides a comprehensive assessment of AI-ECG models for LVSD detection. We evaluate their internal and external performance and examine the transparency of

model development and dataset characteristics to identify potential sources of bias. Additionally, we report on the first-in-kind international collaboration for the independent external validation of available models in a single, well characterized cohort, ensuring a more objective assessment of their generalizability and clinical applicability.

Methods

This systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines **(Supplementary Table 1)**. The institutional review board waived the requirement for informed consent, as the study involves secondary analysis of pre-existing data.

Search Strategy

The Ovid MEDLINE(R) electronic database was systematically searched in October 2023 to identify articles published since January 2010. The employed search strategy was designed to identify studies reporting on the development and/or external validation of AI-ECG models for LVSD detection (**Supplementary Table 2)**. Identified studies were imported into Rayyan for screening and deduplication.[12]

Study selection criteria

All articles describing the development and/or external validation of deep learning (DL) models using ECG data, hereafter referred to as AI-ECG models, were considered for inclusion. We included both standard 12-lead ECGs and reduced-lead recordings (e.g., single-lead wearable data). Studies analyzing ECG signals exceeding 30 seconds in duration were excluded to differentiate AI-ECG models from those applied to continuous ECG recordings, such as Holter or telemetry recordings. From the eligible studies, we further selected only those that specifically addressed the development and/or external validation of AI-ECG models for the detection of LVSD. A detailed overview of the exclusion criteria is provided in the **Supplementary Methods**.

Screening process

All articles were screened for eligibility by two independent reviewers (PC and MB) based on the predefined study selection criteria using study title and abstracts. Disagreements were resolved through discussion. Articles deemed potentially relevant were retrieved in full-text and subsequently assessed for eligibility by the same reviewers.

Data synthesis and analysis

Both reviewers independently extracted data using a standardized data extraction form to ensure completeness and accuracy. General characteristics of the developed models, including their architecture, training cohort, internal and external validation cohort characteristics, data and model availability, and model performance metrics were extracted. A complete list of the extracted variables is provided in **Supplementary Table 3**. Continuous variables are reported as mean ± standard deviation or median with interquartile range and categorical variables were presented as counts with percentages. Performance of models was categorized into four groups: excellent (area under the receiver operating characteristic curve [AUROC] > 0.9), good (0.8-0.9), moderate (0.7-0.8), and poor (<0.7).

External validation of included LVSD models

To obtain access to existing AI-ECG models, we contacted all the corresponding authors of relevant publications. Initial emails were sent on the 16th of July 2024 outlining our request for independent external validation of their models. The email templates for the initial invitation and follow-up email can be found in the **Supplementary Methods**. Specifically, it was requested that the developed AI-ECG models for LVSD detection were shared for local deployment at the Amsterdam UMC. In cases where no response was received, a follow-up email was sent after two weeks to maximize the likelihood of engagement and ensure a fair opportunity for authors to share their models for independent validation. For all responding groups, local technical deployment and

data and model privacy concerns were discussed in meetings, after which the research groups prepared their models for local deployment. Each research group then independently determined which version of their model they wished to provide for external validation.

External validation cohort

Data for this study were sourced from the Adam CMR database (Amsterdam UMC, The Netherlands), a prospectively maintained registry comprising all consecutive patients who underwent cardiac magnetic resonance imaging (CMR) as part of routine clinical care. The database includes over 130 variables, encompassing patient demographics, clinical characteristics, structured qualitative interpretations, and comprehensive quantitative imaging measurements. All CMR scans were performed between 2015 and 2022, with reports reviewed by expert cardiologists.

From all patients within the Adam CMR database, clinically obtained 12-lead ECGs (Philips PageWriter or GE MUSE) were collected and all ECGs recorded within a 45-day window, spanning 15 days prior to and 30 days after CMR, were included in the external validation cohort. LVEF was recorded as a continuous variable for all patients and used to define the LVEF threshold labels.

The second cohort, referred to as the representative heart failure (HF) cohort, is a subsample of the full CMR cohort constructed to reflect the heart failure prevalence and LVEF distribution typically observed in emergency departments and cardiology outpatient clinics. This sub-cohort consisted of 15% of patients with LVEF <40%, 10% with LVEF between 40–50%, and 75% with LVEF (>50%) serving as controls. Among the controls, CMR findings were either normal or limited to isolated aortic pathology or coronary artery disease. This stratification was performed to enable a more clinically relevant evaluation of model performance in a setting where AI-ECG models are most likely to be applied. 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 this evaluation.

Model performance assessment

All shared models were deployed on the ECGs in the CMR cohort to generate predicted probabilities ranging from 0 to 1. Models were distributed as containerized Docker modules or via repositories containing the full pipeline and model weights. Each module ingested raw signals directly because all signal-processing steps were encapsulated within the module. Model performance was evaluated using 500 bootstrap samples to derive confidence intervals (CI). For the full CMR cohort, data were bootstrapped by selecting a different ECG per patient in each iteration. For the representative HF cohort, a different subset of the full cohort was used in every iteration. Performance metrics included the AUROC, sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV). Thresholds were based on values provided by the participating researchers. The number of ECGs for which models did not return valid predictions, for example due to noise, was determined and reported per model. To ensure a fair comparison across models, sensitivity analysis was conducted using only the subset of ECGs returning valid predictions for each model. Results were stratified based on sex (male and female), age (<60 and ≥60 years), QRS duration (<100ms and ≥100ms), heart rate (>80 bpm and ≤80 bpm), atrial fibrillation status (present or absent), and clinical diagnosis of the CMR, including coronary artery disease, ischemia, dilated cardiomyopathy, hypertrophic cardiomyopathy, myocarditis and other cardiomyopathies. Furthermore, false positive rates were determined for LVEF 40-50% or LVEF >50%.

Threshold Optimization

To assess whether recalibrating the threshold could enhance model performance on the Amsterdam UMC CMR dataset, we optimized the model probability threshold corresponding to

Youden's Index in the Amsterdam UMC dataset. We then calculated sensitivity, specificity, NPV, and PPV based on this optimized threshold.

Correlation between individual models

To evaluate the similarity of the outputs generated by the individual models, we calculated pairwise Spearman's rank correlation coefficients using the predicted probabilities of each model.

Results

Systematic review

A total of 4,321 articles were identified in the search. After removing 70 duplicate records, 4,251 articles remained for title and abstract screening, in which 3,680 articles were excluded **(Figure 1)**. An additional 188 articles were excluded during full-text screening. The most common reasons for exclusion were the use of ECG recordings > 30 seconds (e.g. Holter monitoring or telemetry), beat detection models and the use of manual ECG-feature extraction. Ultimately, only articles describing the derivation and/or validation of AI-ECG for LVSD were selected for analysis. Therefore, all articles (n=348) not primarily trained for LVSD, not describing derivation and/or validation, using multi-modal data or that lacked a clear outcome definition, were excluded, resulting in a total of 35 articles included in the final analysis **(Table 1).**

Of the included articles, 9 (25) reported internal validation performance only, 11 (31%) both internal and external validation and 15 (43%) external validation only. All the external validation efforts were conducted by the developers of the models. Of the included articles describing the development of a model, 10 (50%) originated from Asia, 8 (40%) from North America, and 2 (10%) from Europe. In total, all articles included 51 different AI-ECG models for the prediction of LVSD. All models were variations of convolutional neural networks. Of these models, 48 (94%) utilized direct signal data, and 3 (6%) used an ECG-image as input for the model **(Supplementary Tables 4-7)**.

LVSD definitions

In 34 articles (covering 50 models), LVSD was defined using LVEF measured via transthoracic echocardiography (TTE). The cut-off to define LVSD was set at LVEF <35% for 9 models, LVEF <40% for 31 models, LVEF 40-50% in 2 models, LVEF <50% in 6 models. One article did not describe a TTE based LVEF cut-off to define LVSD but instead used registered disease definitions HF with reduced EF, HF with mildly reduced EF, HF with preserved EF and asymptomatic LVSD.

Reported model performance

An overview of the reported baseline characteristics for the internal validation and external validation cohorts of included studies is presented in **Figure 2A**. For both the internal validation and external validation all models reported the AUROC **(Supplementary Table 6 and 7).** In 12 out of the 20 articles describing the derivation of a model, external validation for that model was reported either within the same article or in a separate article **(Supplementary Table 6).**

Model performance for detecting LVSD

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 the external validation, the performance was excellent in 31 (50%) of the models, good in 28 (45%) and moderate in three (5%), and none were poor, respectively **(Supplementary Table 7).** The difference in internal validation and external validation AUROC can be observed in **Figure 2B**.

Quality and bias assessment

Of the articles reporting on the derivation of their models, nine stratified the results to detect potential bias towards subgroups based on age, sex, ethnicity, comorbidity or ECG characteristics **(Supplementary Table 6).**

According to the PROBAST checklist, among all studies reporting on the derivation of AI-ECG models, 8 models (40%) were rated as having a high risk of bias, and another 8 models

(40%) had an unclear risk of bias **(Supplementary Figure 1)**. Concerns regarding applicability were high in 1 model (5%) and unclear in 2 models (10%) **(Supplementary Figure 2)**. High risk of bias was exclusively attributed to the absence of external validation. Unclear risk of bias was primarily due to insufficient or vague descriptions of the study cohort, or cohort characteristics that did not align with intended real-world applications (e.g., unusually low outcome prevalence). High or unclear concern regarding model applicability stemmed from an undefined or inadequately defined outcome (n=1, 5%), outcome definitions that did not reflect real-world clinical scenarios (n=1, 5%), or participant inclusion criteria that were inconsistent with typical clinical practice (n=1, 5%).

Independent external validation

Model availability

None of the reported models and/or the weights were publicly available for independent external validation. After reaching out to 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.[2,13–15] The models originated from Korea, United States (US), Taiwan and The Netherlands.

The research group identified through Cho et al.[13] shared an algorithm called AiTiALVSD v2.00.00 and is based on a transformer-based architecture and convolutional encoders. It was trained on 498,726 ECGs from 186,889 patients across 16 tertiary hospitals in Korea using an 80–10–10% split.[16] 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. shared an AI-ECG model named ECG Vision. This model is a pretrained version of EfficientNet-B3 that was fine-tuned using a dataset of 116,210 patients from Yale New Haven Hospital in the US using an 85-5-10% split for training, validation and testing. The model inputs are 300x300 ECG images either directly or plotted from the raw signal[2]. LVSD was defined

as LVEF <50% 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 Table 6 and 7).**

Huang et al. shared a ResNet-based model trained on signal data from 380,675 patients from the 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.[15] LVSD was defined as LVEF <40% on TTE performed within two weeks of the ECG. External validation was conducted on data from another hospital in Taiwan, yielding an AUROC of 0.94 for internal validation and 0.95 for external validation **(Supplementary Table 6 and 7).**

The research group identified though the publication by van de Leur et al.[14] shared an unpublished ResNet-based model trained on median beat ECGs from 57,250 patients across two hospitals in the Netherlands, using a 90-10 % split for training and internal validation, and evaluated it on a separate internal cohort of newly referred cardiology outpatients. This is a separate model from the original publication, differing in model architecture, training, and development cohort. The pipelined applied a custom algorithm to construct median beat ECGs, which rejected 6% of the ECGs due to noise. 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, comprising 1,203 patients and 4,737 corresponding ECGs, of which we selected one random ECG per individual every bootstrap iteration. The mean age of the population 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 subselection of patients for which all models provided a valid prediction included 94% of all ECGs. For the representative heart failure (HF) cohort, 554 patients

were included in each bootstrap iteration, comprising 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 3, 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. Excluding the ECGs without a valid prediction in any of the models did not affect performance **(Supplementary Table 10)**.

The AiTiALVSD and ECG Vision slightly underestimated the observed risk, whereas the models CGMH and Utrecht slightly overestimated the observed risk **(Figure 3B)**. Recalibration of the thresholds resulted in small differences in sensitivity and specificity. In three models, sensitivity improved slightly at the expense of a minor reduction in specificity, while in one model, specificity increased slightly with a corresponding decrease in sensitivity **(Supplementary Table 11)**.

Stratification by age and sex did not reveal potential bias **(Figure 4, Supplementary Table 8 and 9)**. All four models showed slightly reduced performance in ECGs with QRS width >100ms and in patients with atrial fibrillation. Additionally, AUROCs were generally lower in cardiomyopathy subgroups. Despite these variations, the models maintained good-to-excellent performance across all subgroups **(Figure 4)**. The percentage of false positives ranged from 53% to 65% in patients with LVEF between 40 and 50% and from 12% to 39% in patient with LVEF >50% for the included models **(Supplementary Table 12)**.

Correlation Between Individual Models

Pairwise Spearman's correlation coefficients of the continuous model outputs showed moderate to high correlations across all models (**Figure 3C)**. The strongest correlation was observed between the CGMH model and AiTiALVSD (ρ = 0.86). Correlations with the Utrecht model were generally lower (ρ range 0.69–0.73).

Discussion

We are the first to introduce a novel framework for the external validation of AI-ECG. We provide a systematic, head-to-head comparison of all available AI-ECG models for LVSD detection. Although the majority of these models demonstrate outstanding discrimination (AUROC > 0.90), we identified that the overall risk of bias remains substantial, driven by incomplete characterization of training cohorts and an overreliance on internal or developer-led validations. Crucially, of the 51 models we identified, only four were accessible for independent testing, underscoring significant gaps in transparency and external generalizability. The four models that were available for external validation in a single, well characterized dataset exhibited consistent performance despite disparate training populations. Our research revealed that their diagnostic accuracy hinges on edge-case examples, where false positives predominate, and high inter-model correlations suggest that they are capturing similar electrocardiographic signatures of LVSD.

The strong performance of AI-ECG for LVSD detection in the literature supports the potential of these models to screen patients for early diagnosis of LVSD. However, this study identifies significant gaps in the evaluation of AI-ECG models, particularly in evaluating potential biases, which may limit trustworthiness and consequently impede clinical adoption. External validation in all reviewed articles was conducted solely by the original developers, increasing the risk of selective data inclusion and biased reporting. Despite their promising performance, most models thus lack alignment with the FUTURE-AI guidelines,[6] particularly regarding transparency, reproducibility, and open accessibility. To address this, corresponding authors of all included articles were invited to share their models for independent validation after which four research groups agreed to participate and shared their models.

The independent external validation of the four models showed consistent performance, with modest degradation when comparing the lower-complexity (representative HF) to the higher-complexity (full CMR) cohort. Notably, this external performance remained consistent despite three of the models being trained on geographically distinct cohorts (Korea, US, and Taiwan). Moreover, the continuous probabilities of these models showed high pairwise correlation suggesting that ECG features identified by AI are consistent despite being trained on three continents, supporting generalizability. Additionally, models trained on raw ECG signals or ECG images performed similarly well, aligning with findings from recent literature.[17] It is noteworthy that all included AI-ECG models were originally developed and validated using TTE-derived LVEF as ground truth, whereas our independent validation employed CMR, the reference standard for LVEF assessment known for its superior accuracy and reproducibility. Despite the known difference between these modalities, where TTE tends to underestimate LVEF compared to CMR, the models maintained strong performance, underscoring their robustness and potential clinical utility.[18,19] The slight performance degradation observed in the full CMR cohort prompts careful consideration and potential model recalibration when deployed in highly complex patient populations.

Despite the promising results in the literature and in our independent validation, it is concerning that only a small portion of the identified models were available for external validation. With the recent FDA approval of AI-ECG models,[20] marking their potential transition into clinical practice, it is essential to evaluate the models performance while limiting potential bias. Selective sharing creates a false sense of promise, as researchers with high confidence in their models are more likely to share them, potentially inflating performance.

Our findings highlight how validation cohort characteristics can affect model performance. AUROC differences of ~0.04 between low-complexity (representative HF cohort) and high-complexity (CMR referral) populations illustrate this dependency. Many published models lack clear reporting on cohort composition, raising concerns about bias and limiting applicability.[21] Subgroup variation, particularly in patients with broad QRS or atrial fibrillation, emphasizes the

need for transparent population descriptors. Specific patient characteristics may necessitate model recalibration for safe clinical use.

The slightly lower AUROC observed in the full cohort might reflect the inherent challenge of distinguishing between cases with overlapping, uncommon or specific cardiac diseases, such as patients with subclinical or borderline LVSD and diagnosed with cardiomyopathies with preserved ejection fraction. Previous studies have shown that false positive results in LVSD screening are associated with underlying cardiac conditions or the development of LVSD in the near future.[8,22,23] In the current study, most false positives had an LVEF between 40 and 50% (based on CMR measurement), indicating borderline cardiac dysfunction. These results suggest that, while the models may struggle to distinguish between borderline and definitive LVSD, a positive screen still warrants further clinical evaluation. Such low-threshold, scalable AI-ECG screening tools could play a critical role in the early detection of reduced LVEF, enabling timely initiation of therapy and improving outcomes, particularly in settings with limited access to echocardiography, such as primary care or low-resource environments.

Independent validation is essential to prevent overestimation of the generalizability of AI-ECG models, emphasizing the need for fair and unbiased evaluation. In this study, the well-characterized external validation cohort functioned as a standardized evaluation framework for AI-ECG models. By providing a consistent benchmark for model assessment, independent validation in such a clearing house cohort enables objective comparison across models, mitigating concerns of selective reporting and ensuring a fair assessment of generalizability. Overall, when considering the FUTURE-AI guidelines, key principles as fairness, transparency, robustness and generalizability are addressed in our independent external validation, to further promote trustworthiness in the clinical application and adoption of developed AI-ECG models.

Limitations and Future Directions

We included four models for external validation, all of which demonstrated excellent performance. However, this may 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 included in the external validation provided updated versions of their models that differ from those included in the systematic review, precluding direct comparison of their internal and external performance. While the use of an independent validation set reduces the risk of developer bias, the analysis was limited to a single Western European cohort of patients referred for CMR. To mitigate the effect of cohort-specific selection bias, we simulated a population similar to a population seen in emergency departments or outpatient clinics, in which the performance of the models remained consistent, which is common practice in the evaluation of AI-ECG models.

Future research should prioritize collaborative initiatives, model sharing, and independent external validation across geographically and demographically diverse datasets. Additionally, adopting standardized metrics and benchmarks for the description of dataset characteristics will be essential to ensure consistency and comparability in the literature. With these concerted efforts, AI-ECG models can transition from promising research tools to clinically implemented decision-support systems, delivering tangible benefits to patients.

Conclusion

We provide a novel framework for the external validation and head-to-head comparison of AI-ECG models for LVSD detection. The models included in this external validation demonstrated consistently high discriminative performance, highlighting their potential for clinical application across diverse healthcare settings. However, unclear reporting of development and validation cohort characteristics in many studies raises concerns about potential bias, and limited model availability significantly impedes independent validation hindering broader clinical implementation.

Acknowledgments:

Funding: PC is supported by University of Amsterdam Research Priority Agenda Program AI for Heath 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).

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.

Disclosure of interests: 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.

Code availability: Analyses were conducted using python 3.11. For full code availability see repository: <URL AVAILABLE UPON PUBLICATION>.

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. Khunte A, Sangha V, Oikonomou EK, Dhingra LS, Aminorroaya A, Coppi A, Shankar SV, Mortazavi BJ, Bhatt DL, Krumholz HM, Nadkarni GN, Vaid A, Khera R. Automated diagnostic reports from images of electrocardiograms at the Point-of-care. *medRxiv* 2024.

2. Sangha V, Nargesi AA, Dhingra LS, Khunte A, Mortazavi BJ, Ribeiro AH, Banina E, Adeola O, Garg N, Brandt CA, Miller EJ, Ribeiro ALP, Velazquez EJ, Giatti L, Barreto SM, Foppa M, Yuan N, Ouyang D, Krumholz HM, Khera R. Detection of left ventricular systolic dysfunction from electrocardiographic images. *Circulation* 2023;**148**:765–777.

3. Noseworthy PA, Attia ZI, Behnken EM, Giblon RE, Bews KA, Liu S, Gosse TA, Linn ZD, Deng Y, Yin J, Gersh BJ, Graff-Radford J, Rabinstein AA, Siontis KC, Friedman PA, Yao X. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. *Lancet* 2022;**400**:1206–1212.

4. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Friedman PA. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* 2019;**25**:70–74.

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

6. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, Weicken E, Asselbergs FW, Prior F, Collins GS, Kaissis G, Tsakou G, Buvat I, Kalpathy-Cramer J, Mongan J, Schnabel JA, Kushibar K, Riklund K, Marias K, Amugongo LM, Fromont LA, Maier-Hein L, Cerdá-Alberich L, Martí-Bonmatí L, Cardoso MJ, Bobowicz M, Shabani M, Tsiknakis M, Zuluaga MA, Fritzsche M-C, Camacho M, Linguraru MG, Wenzel M, De Bruijne M, Tolsgaard MG, Goisauf M, Cano Abadía M, Papanikolaou N, Lazrak N, Pujol O, Osuala R, Napel S,

Colantonio S, Joshi S, Klein S, Aussó S, Rogers WA, Salahuddin Z, Starmans MPA, FUTURE-AI Consortium. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 2025;**388**:e081554.

7. 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.

8. Sangha V, Dhingra LS, Oikonomou E, Aminorroaya A, Sikand NV, Sen S, Krumholz HM, Khera R. Identification of hypertrophic cardiomyopathy on electrocardiographic images with deep learning. *medRxiv* 2023.

9. Ledwidge M, Gallagher J, Conlon C, Tallon E, O'Connell E, Dawkins I, Watson C, O'Hanlon R, Bermingham M, Patle A, Badabhagni MR, Murtagh G, Voon V, Tilson L, Barry M, McDonald L, Maurer B, McDonald K. Natriuretic peptide-based screening and collaborative care for heart failure: the STOP-HF randomized trial. *JAMA* 2013;**310**.

10. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failurehttps://www.escardio.org/Guidelines/Clinical-Practice-Guidelines/Acute-and-Chronic-Heart-Failure (January 21, 2025)

11. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, Deswal A, Drazner MH, Dunlay SM, Evers LR, Fang JC, Fedson SE, Fonarow GC, Hayek SS, Hernandez AF, Khazanie P, Kittleson MM, Lee CS, Link MS, Milano CA, Nnacheta LC, Sandhu AT, Stevenson LW, Vardeny O, Vest AR, Yancy CW, ACC/AHA Joint Committee Members. 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.

12. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev* 2016;**5**:210.

13. Cho J, Lee B, Kwon J-M, Lee Y, Park H, Oh B-H, Jeon K-H, Park J, Kim K-H. Artificial intelligence algorithm for screening heart failure with reduced ejection fraction using electrocardiography. *ASAIO J* 2021;**67**:314–321.

14. Leur RR van de, Bos MN, Taha K, Sammani A, Yeung MW, Duijvenboden S van, Lambiase PD, Hassink RJ, Harst P van der, Doevendans PA, Gupta DK, Es R van. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *Eur Heart J Digit Health* 2022;**3**:390–404.

15. Huang Y-C, Hsu Y-C, Liu Z-Y, Lin C-H, Tsai R, Chen J-S, Chang P-C, Liu H-T, Lee W-C, Wo H-T, Chou C-C, Wang C-C, Wen M-S, Kuo C-F. Artificial intelligence-enabled electrocardiographic screening for left ventricular systolic dysfunction and mortality risk prediction. *Front Cardiovasc Med* 2023;**10**:1070641.

16. Jung YM, Kang S, Son JM, Lee HS, Han GI, Yoo A-H, Kwon J-M, Park C-W, Park JS, Jun JK, Lee MS, Lee SM. Electrocardiogram-based deep learning model to screen peripartum cardiomyopathy. *Am J Obstet Gynecol MFM* 2023;**5**:101184.

17. Sau A, Zeidaabadi B, Patlatzoglou K, Pastika L, Ribeiro AH, Sabino E, Peters NS, Ribeiro ALP, Kramer DB, Waks JW, Ng FS. A comparison of artificial intelligence-enhanced electrocardiography approaches for prediction of time-to-mortality using electrocardiogram images. *Eur Heart J Digit Health* 2024:ztae090.

18. Haan S de, Boer K de, Commandeur J, Beek AM, Rossum AC van, Allaart CP. Assessment of left ventricular ejection fraction in patients eligible for ICD therapy: Discrepancy between cardiac magnetic resonance imaging and 2D echocardiography. *Neth Heart J* 2014;**22**:449–455.

19. Gardner BI, Bingham SE, Allen MR, Blatter DD, Anderson JL. Cardiac magnetic resonance versus transthoracic echocardiography for the assessment of cardiac volumes and regional function after myocardial infarction: an intrasubject comparison using simultaneous intrasubject recordings. *Cardiovasc Ultrasound* 2009;**7**:38.

20. Anumana ECG-AI LEFhttps://anumana.ai/ecg-ai-lef (February 11, 2025)

21. Galanty M, Luitse D, Noteboom SH, Croon P, Vlaar AP, Poell T, Sanchez CI, Blanke T, Išgum I. 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.

22. Dhingra LS, Aminorroaya A, Sangha V, Pedroso AF, Asselbergs FW, Brant LCC, Barreto SM, Ribeiro ALP, Krumholz HM, Oikonomou EK, Khera R. Heart failure risk stratification using artificial intelligence applied to electrocardiogram images: a multinational study. *Eur Heart J* 2025;**46**:1044–1053.

23. Croon PM, Dhingra LS, Biswas D, Oikonomou EK, Khera R. Specificity of Artificial intelligence-enhanced electrocardiography for cardiovascular diagnosis and risk prediction. *medRxiv*.

24. Bergquist JA, Zenger B, Brundage J, MacLeod RS, Bunch TJ, Shah R, Ye X, Lyons A, Torre M, Ranjan R, Tasdizen T, Steinberg BA. Performance of off-the-shelf machine learning architectures and biases in low left ventricular ejection fraction detection. *Heart Rhythm O2* 2024;**5**:644–654.

25. Khunte A, Sangha V, Oikonomou EK, Dhingra LS, Aminorroaya A, Mortazavi BJ, Coppi A, Brandt CA, Krumholz HM, Khera R. Detection of left ventricular systolic dysfunction from single-lead electrocardiography adapted for portable and wearable devices. *NPJ Digit Med* 2023;**6**:124.

26. Vaid A, Jiang J, Sawant A, Lerakis S, Argulian E, Ahuja Y, Lampert J, Charney A, Greenspan H, Narula J, Glicksberg B, Nadkarni GN. A foundational vision transformer improves diagnostic performance for electrocardiograms. *NPJ Digit Med* 2023;**6**:108.

27. Honarvar H, Agarwal C, Somani S, Vaid A, Lampert J, Wanyan T, Reddy VY, Nadkarni GN, Miotto R, Zitnik M, Wang F, Glicksberg BS. 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.

28. Chen H-Y, Lin C-S, Fang W-H, Lee C-C, Ho C-L, Wang C-H, Lin C. 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.

29. Golany T, Radinsky K, Kofman N, Litovchik I, Young R, Monayer A, Love I, Tziporin F, Minha I, Yehuda Y, Ziv-Baran T, Fuchs S, Minha S. Physicians and machine-learning algorithm performance in predicting left-ventricular systolic dysfunction from a standard 12-lead-electrocardiogram. *J Clin Med* 2022;**11**:6767.

30. Lee C-H, Liu W-T, Lou Y-S, Lin C-S, Fang W-H, Lee C-C, Ho C-L, Wang C-H, Lin C. Artificial intelligence-enabled electrocardiogram screens low left ventricular ejection fraction with a degree of confidence. *Digit Health* 2022;**8**:20552076221143250.

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

32. Kwon J-M, Jo Y-Y, Lee SY, Kang S, Lim S-Y, Lee MS, Kim K-H. Artificial intelligence-enhanced smartwatch ECG for heart failure-reduced ejection fraction detection by generating 12-lead ECG. *Diagnostics (Basel)* 2022;**12**:654.

33. Surendra K, Nürnberg S, Bremer JP, Knorr MS, Ückert F, Wenzel JP, Bei der Kellen R, Westermann D, Schnabel RB, Twerenbold R, Magnussen C, Kirchhof P, Blankenberg S, Neumann J, Schrage B. Pragmatic screening for heart failure in the general population using an electrocardiogram-based neural network. *ESC Heart Fail* 2023;**10**:975–984.

34. Vaid A, Johnson KW, Badgeley MA, Somani SS, Bicak M, Landi I, Russak A, Zhao S, Levin MA, Freeman RS, Charney AW, Kukar A, Kim B, Danilov T, Lerakis S, Argulian E, Narula J, Nadkarni GN, Glicksberg BS. Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. *JACC Cardiovasc Imaging* 2022;**15**:395–410.

35. Katsushika S, Kodera S, Nakamoto M, Ninomiya K, Inoue S, Sawano S, Kakuda N, Takiguchi H, Shinohara H, Matsuoka R, Ieki H, Higashikuni Y, Nakanishi K, Nakao T, Seki T, Takeda N, Fujiu K, Daimon M, Akazawa H, Morita H, Komuro I. The effectiveness of a deep learning model to detect left ventricular systolic dysfunction from electrocardiograms. *Int Heart J* 2021;**62**:1332–1341.

36. 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.

37. Sun J-Y, Qiu Y, Guo H-C, Hua Y, Shao B, Qiao Y-C, Guo J, Ding H-L, Zhang Z-Y, Miao L-F, Wang N, Zhang Y-M, Chen Y, Lu J, Dai M, Zhang C-Y, Wang R-X. A method to screen left ventricular dysfunction through ECG based on convolutional neural network. *J Cardiovasc Electrophysiol* 2021;**32**:1095–1102.

38. Attia ZI, Harmon DM, Dugan J, Manka L, Lopez-Jimenez F, Lerman A, Siontis KC, Noseworthy PA, Yao X, Klavetter EW, Halamka JD, Asirvatham SJ, Khan R, Carter RE,

Leibovich BC, Friedman PA. Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction. *Nat Med* 2022;**28**:2497–2503.

39. Harmon DM, Adedinsewo D, Van't Hof JR, Johnson M, Hayes SN, Lopez-Jimenez F, Jones C, Attia ZI, Friedman PA, Patten CA, Cooper LA, Brewer LC. 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.

40. Harmon DM, Carter RE, Cohen-Shelly M, Svatikova A, Adedinsewo DA, Noseworthy PA, Kapa S, Lopez-Jimenez F, Friedman PA, Attia ZI. 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.

41. Attia ZI, Dugan J, Rideout A, Maidens JN, Venkatraman S, Guo L, Noseworthy PA, Pellikka PA, Pham SL, Kapa S, Friedman PA, Lopez-Jimenez F. 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.

42. Attia IZ, Tseng AS, Benavente ED, Medina-Inojosa JR, Clark TG, Malyutina S, Kapa S, Schirmer H, Kudryavtsev AV, Noseworthy PA, Carter RE, Ryabikov A, Perel P, Friedman PA, Leon DA, Lopez-Jimenez F. External validation of a deep learning electrocardiogram algorithm to detect ventricular dysfunction. *Int J Cardiol* 2021;**329**:130–135.

43. Attia ZI, Kapa S, Yao X, Lopez-Jimenez F, Mohan TL, Pellikka PA, Carter RE, Shah ND, Friedman PA, Noseworthy PA. 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.

44. Klein CJ, Ozcan I, Attia ZI, Cohen-Shelly M, Lerman A, Medina-Inojosa JR, Lopez-Jimenez F, Friedman PA, Milone M, Shelly S. 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.

45. Bachtiger P, Petri CF, Scott FE, Ri Park S, Kelshiker MA, Sahemey HK, Dumea B, Alquero R, Padam PS, Hatrick IR, Ali A, Ribeiro M, Cheung W-S, Bual N, Rana B, Shun-Shin M, Kramer DB, Fragoyannis A, Keene D, Plymen CM, Peters NS. 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.

46. Kashou AH, Medina-Inojosa JR, Noseworthy PA, Rodeheffer RJ, Lopez-Jimenez F, Attia IZ, Kapa S, Scott CG, Lee AT, Friedman PA, McKie PM. Artificial intelligence-augmented electrocardiogram detection of left ventricular systolic dysfunction in the general population. *Mayo Clin Proc* 2021;**96**:2576–2586.

47. Brito BO de F, Attia ZI, Martins LNA, Perel P, Nunes MCP, Sabino EC, Cardoso CS, Ferreira AM, Gomes PR, Luiz Pinho Ribeiro A, Lopez-Jimenez F. 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.

48. Kashou AH, Noseworthy PA, Lopez-Jimenez F, Attia ZI, Kapa S, Friedman PA, Jentzer JC. 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.

49. Jentzer JC, Kashou AH, Attia ZI, Lopez-Jimenez F, Kapa S, Friedman PA, Noseworthy PA. Left ventricular systolic dysfunction identification using artificial intelligence-augmented electrocardiogram in cardiac intensive care unit patients. *Int J Cardiol* 2021;**326**:114–123.

50. Adedinsewo D, Carter RE, Attia Z, Johnson P, Kashou AH, Dugan JL, Albus M, Sheele JM, Bellolio F, Friedman PA, Lopez-Jimenez F, Noseworthy PA. 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.

51. Noseworthy PA, Attia ZI, Brewer LC, Hayes SN, Yao X, Kapa S, Friedman PA, Lopez-Jimenez F. 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.

52. Vaid A, Jiang JJ, Sawant A, Singh K, Kovatch P, Charney AW, Charytan DM, Divers J, Glicksberg BS, Chan L, Nadkarni GN. Automated determination of left ventricular function using electrocardiogram data in patients on maintenance hemodialysis. *Clin J Am Soc Nephrol* 2022;**17**:1017–1025.

1 **TABLES AND FIGURES**

2 **Table 1:** Overview of all included articles

First Author	Year of publication	Outcome label	Validation
Bergquist et al.[24]	2023	LVEF <40%	Internal validation
Khunte et al.[25]	2023	LVEF < 40%	Internal validation
Vaid et al.[26]	2023	LVEF < 40%	Internal and external validation
Honarvar et al.[27]	2022	LVEF <35%	Internal validation
Huang et al.[15]	2023	LVEF < 40%	Internal and external validation
Chen et al.[28]	2022	LVEF <50%, <50%	Internal and external validation
Golany et al.[29]	2022	LVEF <35%	Internal validation
Yagi et al.[28]	2022	LVEF <40%	Internal and external validation
Lee et al.[30]	2022	LVEF <40%	Internal and external validation
Sangha et al.[2]	2023	LVEF <40%	Internal and external validation
Chen et al.[31]	2022	LVEF <35%,	Internal and external validation
Kwon et al.[32]	2022	LVEF <40%	Internal and external validation
Surendra et al.[33]	2023	Asymptomatic LVSD, HFrEF, HFmrEF, HFpEF	Internal validation
Vaid et al.[34]	2022	LVEF <35%	Internal and external validation
Katsushika et al.[35]	2021	LVEF <40%	Internal validation
Cho et al.[13]	2021	LVEF <40%	Internal and external validation
Choi et al.[36]	2022	LVEF <40%	External validation model Cho (2021)
Sun et al.[37]	2021	LVEF <50%	internal validation
Attia et al.[36]	2019	LVEF <35%	internal validation
Attia et al.[38]	2022	LVEF < 40%	External validation model Attia (2019)

Harmon et al.[39]	2022	LVEF < 50%, <40%	External validation model Attia (2019)
Harmon et al.[40]	2022	LVEF <40%	External validation model Attia (2019)
Attia et al.[41]	2022	LVEF < 35%, <40%, <50%	External validation model Attia (2019)
Attia et al.[42]	2021	LVEF <35%	External validation model Attia (2019)
Attia et al.[43]	2019	LVEF < 35%	External validation model Attia (2019)
Klein et al.[44]	2022	LVEF <35%	External validation model Attia (2019)
Bachtiger et al.[45]	2021	LVEF < 40%	External validation model Attia (2019)
Kashou et al.[46]	2021	LVEF <35%	External validation model Attia (2019)
Brito et al.[47]	2021	LVEF <40%	External validation model Attia (2019)
Kashou et al.[48]	2021	LVEF 35%	External validation model Attia (2019)
Jentzer et al.[49]	2021	LVEF <35%	External validation model Attia (2019)
Adedinsewo et al.[50]	2020	LVEF <50%	External validation model Attia (2019)
Noseworthy et al.[51]	2020	LVEF <35%	External validation model Attia (2019)
van de Leur et al.[14]	2022	LVEF <40%	Internal and external validation
Vaid et al.[52]	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.

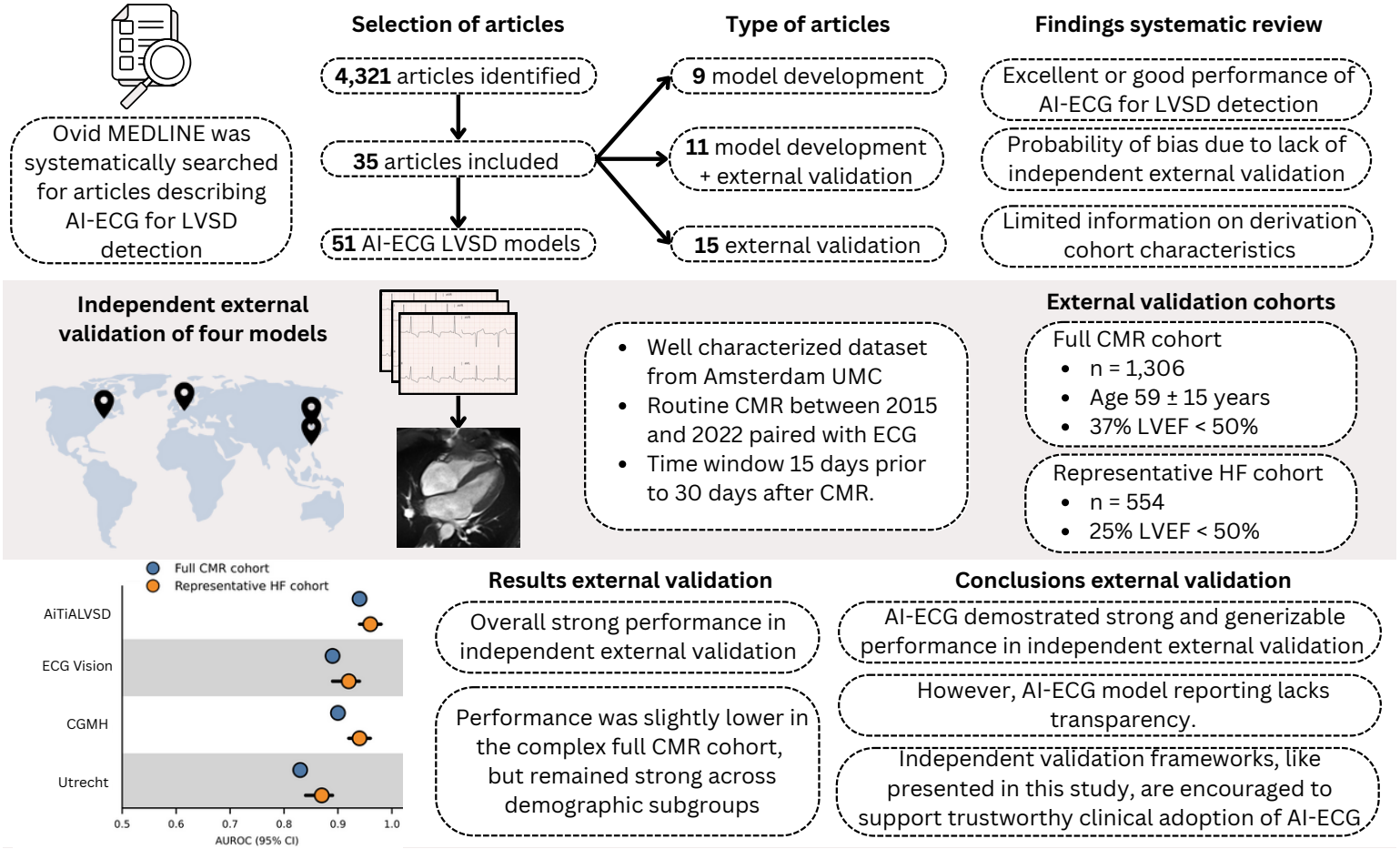
1 **Table 2:** Baseline characteristics of the Amsterdam UMC CMR-cohort.

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)

2

3 Diagnosis are derived from the conclusion of the cardiac magnetic resonance imaging study
4 reported by an expert cardiologist. Abbreviations: n, number of samples; SD, Standard
5 Deviation; LVEF, Left Ventricular Ejection Fraction.

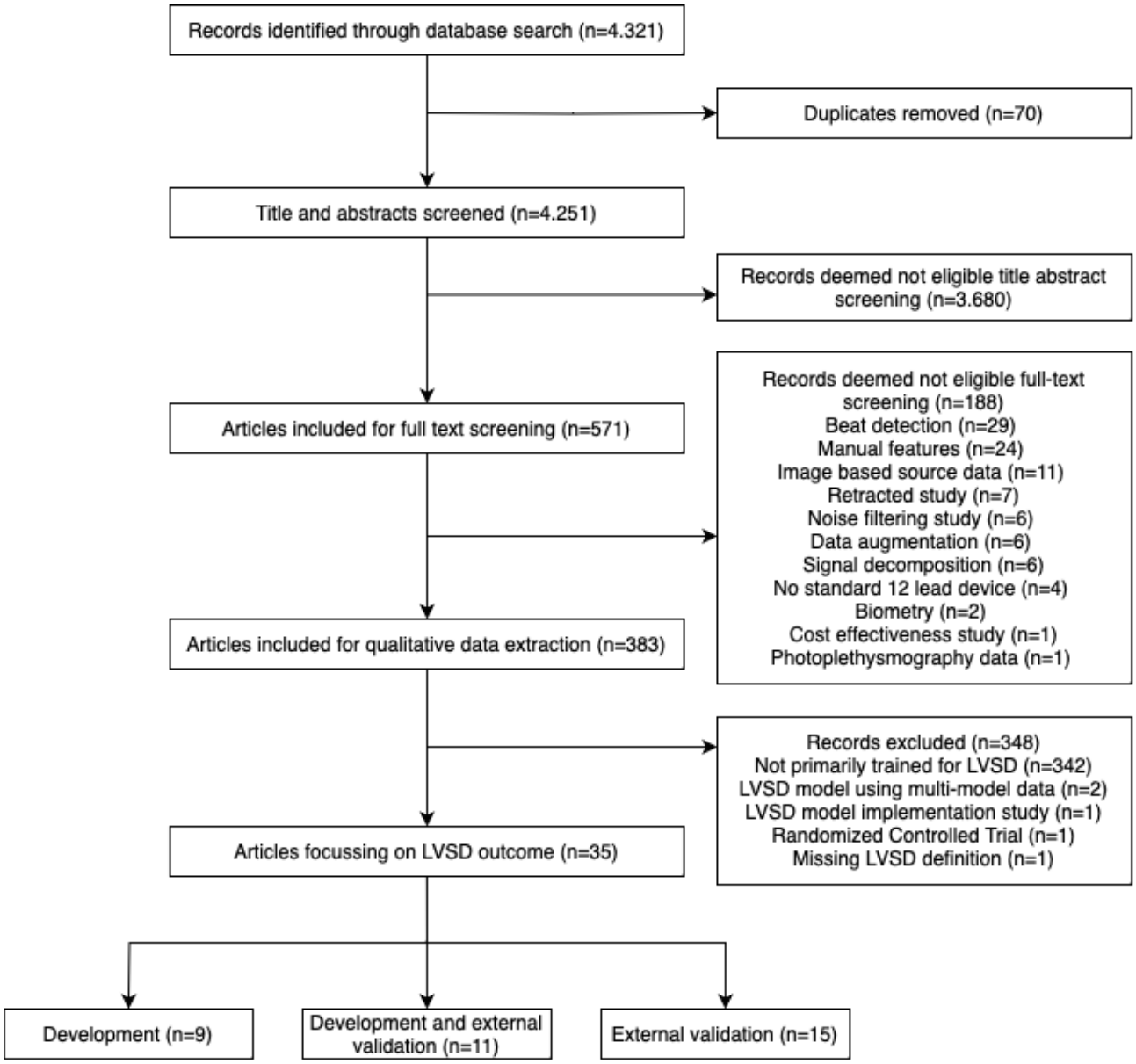
Graphical abstract



Abbreviations: AI, artificial intelligence; ECG, electrocardiogram; LVSD, left ventricular systolic disease; UMC, university medical centers; CMR, cardiac magnetic resonance: LVEF, left ventricular ejection fraction; HF, heart failure; CI, confidence interval.

1

2 **Figure 1:** Flowchart of Study Selection for AI-ECG Models in LVSD Detection



3

4 Abbreviations: LVSD, Left Ventricular Systolic Dysfunction.

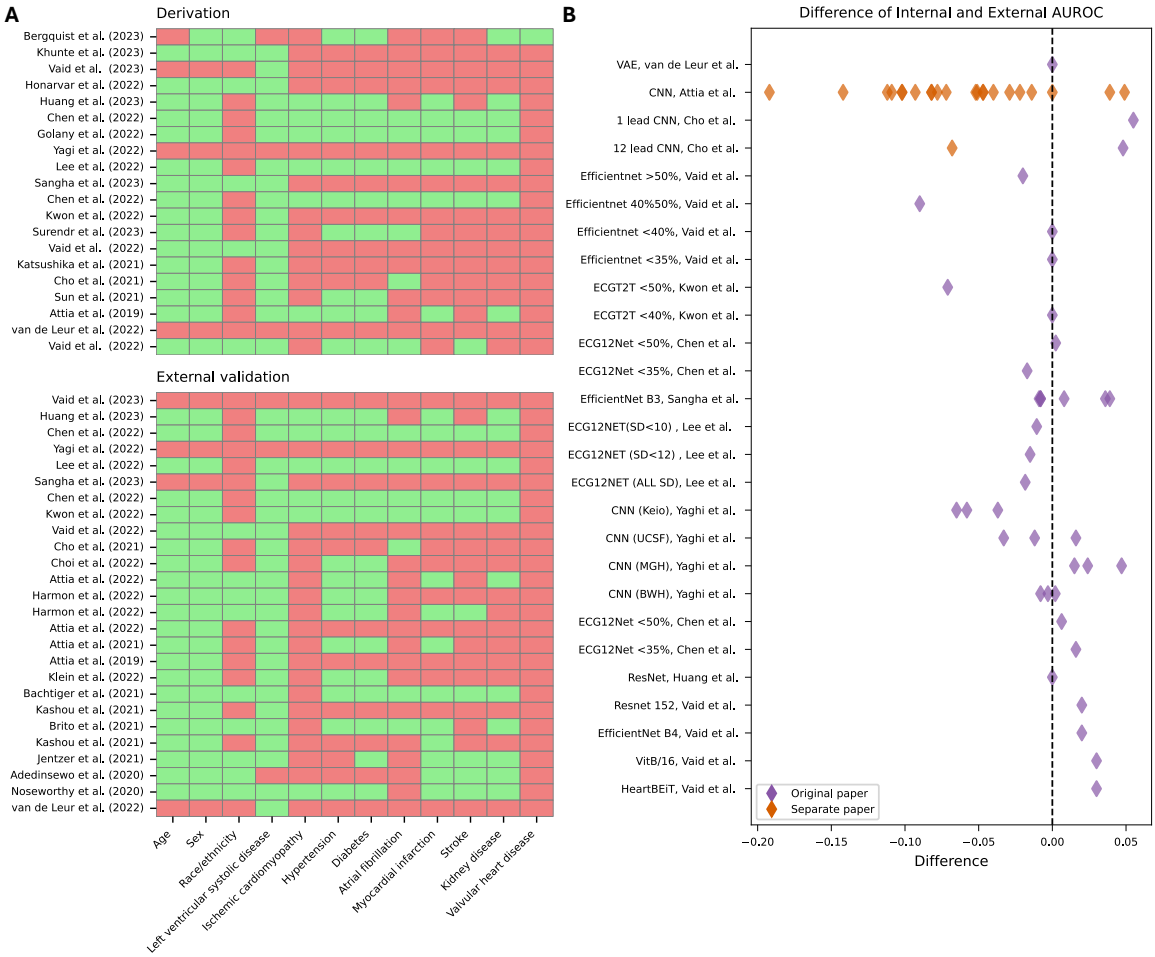
5

6

7

8

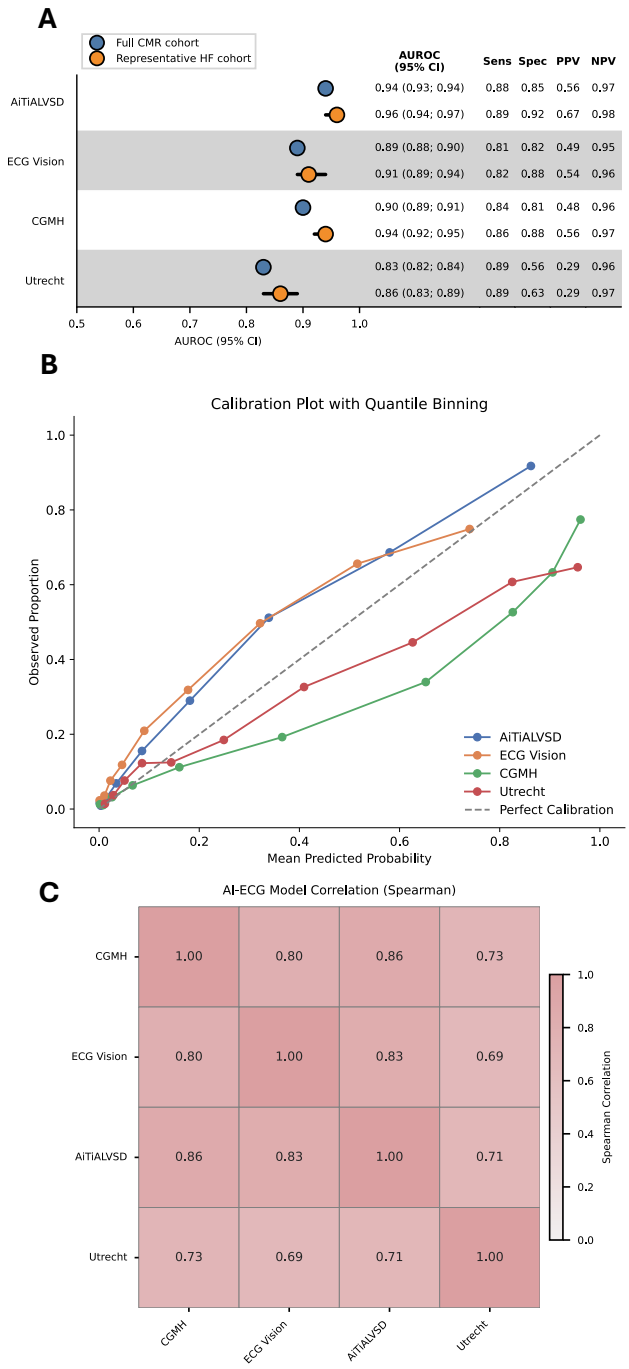
Figure 2: Reported baseline characteristics, internal and external performance for included models.



NB: Panel A: This heatmaps show baseline characteristics reported in derivation (top) and external validation (bottom) studies. Each row represents a study, and each column represents a specific characteristic (e.g., age, sex, comorbidities). Green indicates that the characteristic was reported, while red indicates it was not. Panel B: 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 indicate external validation conducted in separate papers. All values are referenced to internal AUROC, thus a value higher than zero indicate a higher internal AUROC compared to AUROC obtained during

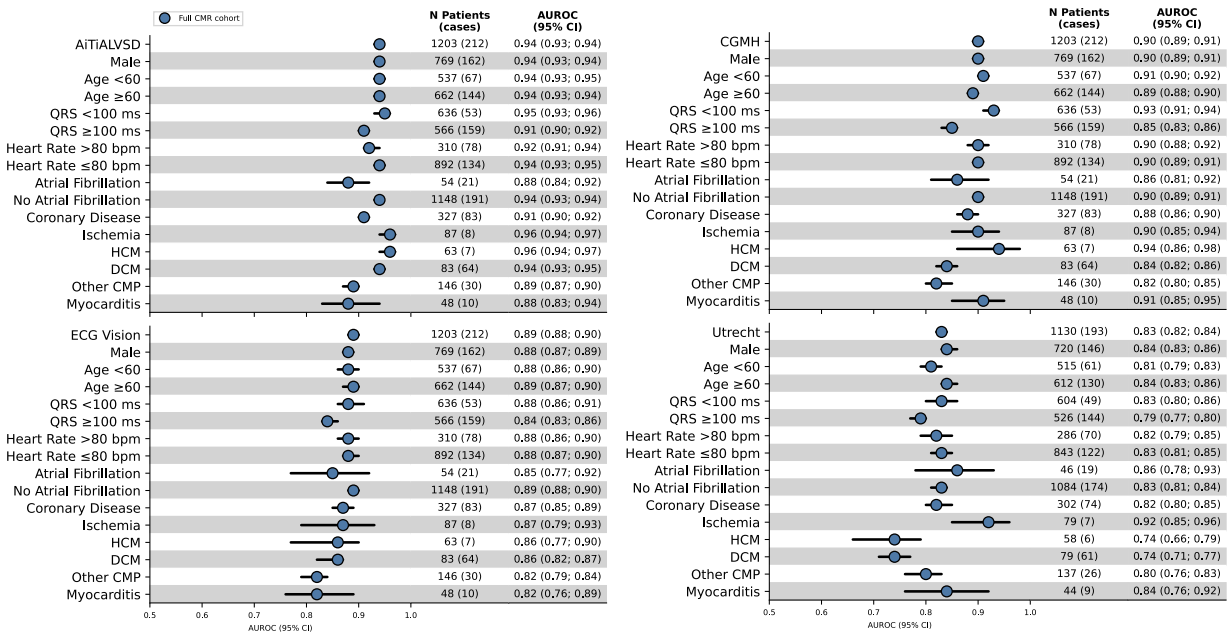
external validation. Abbreviations: AUROC, Area Under the Receiver Operating Curve; CMR, Cardiac Magnetic Resonance imaging; HF, Heart failure.

Figure 3: Performance and calibration of included models for external validation



NB. **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 corresponding 95% confidence intervals (CI). **Panel B:** Calibration plot for the external validation. Mean predicted probabilities versus observed proportions, visualized using quantile binning (10 bins). The "Perfect Calibration" line represents ideal model calibration. **Panel C:** The heatmap displays the pairwise Pearson correlation coefficient 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, CGMH, Chang Gung Memorial Hospital.

Figure 4: Subgroup analysis of AUROC values for external validation of models.



NB: Performance in the full CMR cohort across various patient subgroups is displayed, including sex, age, QRS duration, heart rate, atrial fibrillation status, coronary disease, cardiomyopathy types, and myocarditis. The total number of patients is presented, as well as the number of cases with LVSD in the selected category (n patients (cases EF<40%)). CMR; Cardiac

Magnetic Resonance imaging, AUROC values are presented with 95% confidence intervals, providing insights into model performance within specific populations. Abbreviations: AUROC, Area Under the Receiver Operating Characteristic Curve; HCM, Hypertrophic Cardiomyopathy; DCM, Dilated Cardiomyopathy; CMP, Cardiomyopathy; CGMH, Chang Gung Memorial Hospital.