

Artificial intelligence-based identification of left ventricular systolic dysfunction from 12-lead electrocardiograms: external validation and advanced application of an existing model

Sebastian König ()^{1,2,*†}, Sven Hohenstein^{2,†}, Anne Nitsche ()², Vincent Pellissier ()², Johannes Leiner ()^{1,2}, Lars Stellmacher^{1,2}, Gerhard Hindricks¹, and Andreas Bollmann ()^{1,2}

¹Department of Electrophysiology, Heart Center Leipzig at University of Leipzig, Strümpellstr. 39, 04289 Leipzig, Germany; and ²Helios Health Institute, Real World Evidence & Health Technology Assessment, Schwanebecker Chaussee 50, 13125 Berlin, Germany

Received 7 August 2023; revised 6 December 2023; accepted 14 December 2023; online publish-ahead-of-print 20 December 2023

Aims	The diagnostic application of artificial intelligence (AI)-based models to detect cardiovascular diseases from electrocardio- grams (ECGs) evolves, and promising results were reported. However, external validation is not available for all published algorithms. The aim of this study was to validate an existing algorithm for the detection of left ventricular systolic dysfunction (LVSD) from 12-lead ECGs.
Methods and results	Patients with digitalized data pairs of 12-lead ECGs and echocardiography (at intervals of \leq 7 days) were retrospectively selected from the Heart Center Leipzig ECG and electronic medical records databases. A previously developed Al-based model was applied to ECGs and calculated probabilities for LVSD. The area under the receiver operating characteristic curve (AUROC) was computed overall and in cohorts stratified for baseline and ECG characteristics. Repeated echocardiography studies recorded \geq 3 months after index diagnostics were used for follow-up (FU) analysis. At baseline, 42 291 ECG-echocardiography pairs were analysed, and AUROC for LVSD detection was 0.88. Sensitivity and specificity were 82% and 77% for the optimal LVSD probability cut-off based on Youden's J. AUROCs were lower in ECG subgroups with tachycardia, atrial fibrillation, and wide QRS complexes. In patients without LVSD at baseline and available FU, model-generated high probability for LVSD was associated with a four-fold increased risk of developing LVSD during FU.
Conclusion	We provide the external validation of an existing Al-based ECG-analysing model for the detection of LVSD with robust performance metrics. The association of false positive LVSD screenings at baseline with a deterioration of ventricular function during FU deserves a further evaluation in prospective trials.

[†]The first two authors contributed equally to the study.

^{*} Corresponding author. Tel: +49 341 865 252 613, Fax: +49 341 865 1460, Email: sebastian.koenig@helios-gesundheit.de

[©] The Author(s) 2023. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Graphical Abstract



Keywords

Artificial intelligence • Electrocardiogram • Left ventricular systolic dysfunction • Heart failure

Introduction

Due to an increasing prevalence over the last years, chronic heart failure (HF) has become one of the most relevant cardiovascular diseases with respect to its medical and socioeconomic impact on health care.^{1,2} An improvement of screening strategies to detect early disease stages and asymptomatic patients has been proposed as an important goal especially in patients with HF and left ventricular systolic dysfunction (LVSD) in order to provide therapies that were shown to improve patients' outcomes.³ To date, neither clinical models nor diagnostic tests are established as regular screening tools for asymptomatic LVSD in Europe. Even the measurement of N-terminal pro-brain natriuretic peptide (NT-proBNP) levels was shown to be of only modest sensitivity.⁴ The advanced analysis of standard surface 12-lead electrocardiograms (ECGs) augmented by artificial intelligence (AI)-based algorithms was introduced as a non-invasive alternative with promising performance measures for the detection of LVSD and superior discriminatory power when compared to NT-proBNP measurement.⁵⁻⁹ However, independent external validation is not available for all of the published models.^{10,11} A high variance of performance results for external validation studies reflects both the necessity and difficulty of such analyses, which are influenced by the setting of model application.¹² Validating an algorithm in an unrelated patient population by a different group of researchers is necessitated to prove its reliability and overcome possible systematic bias.¹³ Yagi et al.¹⁴ published a free-to-use algorithm for the detection of LVSD in 2022, but relevant indicators of model performance as well as values needed for the interpretation and external application were not provided. The model has been developed and tested in populations from North America and Japan with unknown baseline characteristics including an unpublished prevalence of LVSD. Since both patientrelated factors and disease prevalence likely influence the prediction of such models, a validation in a European population was considered relevant prior to an utilization in clinical practice.^{15,16} Aims of this study were, therefore, to externally validate an existing algorithm for the detection of LVSD, to describe relevant values and performance metrics in detail, to test strategies for further model improvement, and to investigate whether the model output also predicts future development of LVSD in patients with preserved left ventricular ejection fraction (LVEF) at baseline

Methods

We performed a retrospective external validation of a previously presented Al-based algorithm for the detection of LVSD from standard surface 12-lead ECGs.¹⁴ The algorithm was applied to 12-lead ECGs from the monocentric Heart Center Leipzig ECG database recorded between January 2016 and December 2022. Electrocardiograms from both inpatients and patients from outpatient clinics affiliated with the Heart Center Leipzig were included. Only ECGs from patients aged 20 years or older were used according to the selection criteria used in the initial publication. There was no further patient selection (e.g. no exclusion of patients with clinically prevalent HF or known LVSD). Within the Heart Center Leipzig ECG database, ECGs were stored in.xml format with a sampling rate of 500-2000 Hz and a duration of 10.0-20.0 s. All ECGs were down-sampled to 250 Hz and truncated after the first 10 s. Electrocardiograms were written, and automated analysis of digitalized ECG data was performed both by products of Spacelabs Healthcare GmbH (Snoqualmie, WA, USA). Only ECGs with available information on echocardiography-based LVEF were selected for further analyses, and all ECGs of adequate quality based on automated ECG software evaluation were analysed (possible inclusion of multiple ECGs per patient). Echocardiography was performed using a standardized examination protocol and LVEF was assessed either biplane (Simpson's method) or triplane with LVSD being defined as LVEF < 40% according to the original publication.¹⁴ Imaging results were extracted from electronic medical records (EMR) and were considered valid only if performed within 7 days from ECG recording. For all patients with a valid ECG-echocardiography pair at baseline, EMR data were searched for follow-up (FU) echocardiography containing LVEF information that was recorded at least 3 months after the first imaging study. All data were anonymized prior to further analysis. There were no missing data.

Electrocardiogram data were applied to a freely accessible (web interface for model application accessible under http://onebraveideaml.org/) convolutional neural network-based model (time required for processing one ECG dataset including data upload: ~25 s), that is described in more detail in the original publication of Yagi et al.¹⁴ and under at the following URL: https:// github.com/obi-ml-public/ECG-LV-Dysfunction. The model's output is a probability for existing LVSD expressed by a continuous variable between 0 and 1. Artificial intelligence-based ECG analysis was performed both utilizing raw and pre-processed ECG data (https://github.com/PierreElias/ IntroECG), the latter implying the elimination of baseline-shifting and outlier voltages in a subgroup of randomly selected ECGs from our overall cohort as described previously.¹⁷ Considering the use of anonymized clinical routine data, individual informed consent was not obtained. The study was approved by the responsible ethics committee and follows the TRIPOD reporting guidelines for model validation studies.

Model performance was expressed by area under the receiver operating characteristic curve (AUROC, software package used for creation of AUROC graph: https://github.com/overdodactyl/diagnosticSummary/) together with 95% confidence intervals (CIs), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Youden's J was used to define the optimal cut-off for the predicted probability of LVSD, since no cut-off value was published for the application of the original model. We determined the optimal cut-off for maximizing Youden's J statistic via bootstrapping with 1000 iterations.¹⁸ In order to avoid overestimation of performance variables and positive bias, we report quality statistics based on out-of-bag samples together with Cls. With this approach, calculations of cut-off point and performance were done with independent datasets in each bootstrap run. Model performance measures were calculated stratified for age, sex, and different ECG parameters that were evaluated by automatic ECG software analysis (heart rate, rhythm, PQ interval, QRS duration, abnormal repolarization). We also tested for the impact of the number of analysed ECGs per patient and the time between ECG recording and imaging study on model's performance. In patients with available FU echocardiography, the prediction of developing future LVSD in patients with a LVEF \geq 50% at baseline from index ECGs was tested. For this purpose, the model output (probability of LVSD, included as a logarithmic prediction score) together with additional variables (age, sex, baseline LVEF) were integrated into a multivariable logistic regression model.

Results

Of 216 875 assessable ECGs, 42 291 valid ECG-echocardiography pairs from 31 944 individual patients were analysed. A flow diagram presenting the derivation of the final study cohort is provided in Figure 1. Overall, mean age of cases was 66.2 ± 15.1 years and 38.3% were female. Mean LVEF was 53.3 \pm 13.3%, and prevalence of LVSD upon included cases was 14.9%. The AUROC for the original model using non-preprocessed ECG data from our database was 0.88 (95% CI 0.87-0.88; Figure 2). Based on Youden's I, a cut-off of 0.047 (95% CI 0.040-0.064) for the computed probability of LVSD was identified as the best discriminator with a corresponding sensitivity of 82% (95% CI 0.78-0.84), a specificity of 77% (95% CI 0.75-0.81), an accuracy of 78% (95% CI 0.76-0.80), a PPV of 40% (95% CI 0.38-0.43), and a NPV of 96% (95% CI 0.95–0.96) (Figure 3A and B). Numerical performance measures for the ten model output cut-offs with best Youden's | values are provided in the Supplementary material. When altering the definition of LVSD as part of an exploratory analysis, calculated AUROCs were 0.89 (95% CI 0.89-0.89) for a LVEF cut-off of <35%, and 0.84 (95% CI 0.83-0.84) for a LVEF cut-off of <50%. Plotting AUROC values for different LVEF thresholds on the x-axis, a constantly decreasing curve with a climax in the range between an



Figure 1 Flow diagram showing the derivation of the study cohort.



Figure 2 Receiver operating characteristic curve for overall model performance.

LVEF cut-off of 15–30% was observed (*Figure 4*). When stratifying for age, sex, and specific ECG parameters, an inferior model accuracy based on AUROC was found in subgroups of patients aged \geq 80 years as well as ECGs with heart rate \geq 100 per minute, wide QRS complexes, present pacemaker stimulation, and present atrial fibrillation/ atrial flutter. Neither the time between ECG and echocardiography nor the number and selection of analysed ECGs per patient influenced AUROCs relevantly (*Figure 5*). Electrocardiogram pre-processing was performed in 3185 randomly selected ECGs from 3185 patients (mean age 65.6 ± 15.4 years, 37.7% female, prevalence of LVSD: 14.5%). In this subgroup, there was no difference in model performance when using raw (AUROC 0.89, 95% CI 0.87–0.90) or pre-processed ECG data (AUROC 0.89, 95% CI 0.87–0.90).

In 4669 patients (14.6% of the overall cohort), a FU echocardiography was available (median time from baseline to FU echocardiography 343 days, interquartile range 174-539 days). Within patients with LVEF of \geq 50% at baseline (60.7% of patients with available FU data), 4.2% developed LVSD during FU. Applying the Al-based algorithm to corresponding baseline ECGs, a future deterioration of LVEF to <40% was predicted with an AUROC of 0.68 (95% CI 0.63–0.73). The optimal output probability cut-off based on Youden's | was lower (0.033, 95% CI 0.016–0.045) compared to using the model for LVSD detection at baseline with a resulting sensitivity of 58% (95% Cl 0.43–0.68), a specificity of 74% (95% CI 0.60-0.79), a PPV of 9% (95% CI 0.06-0.11), and a NPV of 98% (95% CI 0.97–0.98). An analysis stratifying AUROCs for different patient-related and ECG characteristics is provided in the Supplementary material. Patients with LVEF \geq 50% at baseline and a high (\geq 3.3%) predicted probability of LVSD had a more than four-fold increased risk (HR 4.04, 95% CI 2.84–5.77) for developing LVSD during FU compared to patients with a low model-based LVSD probability (Figure 6). Integrating the logarithmic AI model output with age, sex, and baseline LVEF into a multivariable model, the AUROC for the prediction of LVSD development during FU was improved to 0.75 (95% CI 0.71–0.79). The AI model-based LVSD probability was an independent predictor for future LVEF deterioration to <40%. Results for odds ratios from univariable and multivariable analyses are summed up in Table 1.

Discussion

With this retrospective analysis based on the Heart Center Leipzig ECG database, we provide an external validation of an existing AI-based model for the detection of LVSD from standard surface 12-lead ECGs. Applying the model to a different patient population from different continents by an unrelated research team, we were able to show a good model discrimination based on the AUROC that was comparable to the performance metrics published for the development of the original model.¹⁴ Calculating an optimal cut-off for the model's outcome probability within our specific population generated an excellent NPV with moderate sensitivity and specificity. The necessity to compute individual cut-offs for patient populations with differing characteristics and a different prevalence of the disease to be detected has been highlighted previously in order to improve results.¹⁹ A further optimization of model performance through ECG pre-processing was not possible.

To date, there are several published Al-based algorithms for the detection of LVSD with reported AUROCs with a median of 0.90 and a range from 0.84 to 0.95, which is congruent with our results.^{10–12,20} However, not all of those algorithms were truly validated in an external cohort with a differing composition of included subjects when compared to the derivation cohort. This is reflected by a high variability of results for existing external validations. One study showed a decrease in AUROC from 0.93 (model development) to 0.82 (external validation), while others reported even higher AUROCs within their validation cohorts when compared to the model development performance metrics.^{19,21,22} Furthermore, model architecture including used weights and output cut-offs were not made publicly available for all presented models, which hinders comprehensive and independent external validation as well as further model application in clinical practice. Several points have to be considered for the interpretation of those results, like the study-specific prevalence of LVSD, the clinical setting of ECG and echocardiography acquisition, the differing quality and device-based pre-processing of ECGs, a differing assessment and definition of LVSD (quality of echocardiography imaging), and others. One major influencing factor refers to the selection of the patient cohort used for external validation. On the one hand, some AI models were applied to a population that was not related to the derivation cohort at all, whereas, in other studies, validation was performed in relatively homogenous populations being divided only per referring hospitals that were in close proximity to each other.^{19,21–23} Even though there is no further information available on baseline characteristics of the patient cohort used for model development within the original publication, the two cohorts are not related at all and inpatient databases from different continents were used to create them.¹⁴

With regard to reported sensitivity, specificity, PPV, and NPV, our findings are mid-range when compared to previously published data.^{5,6,9,24-26} The model has a particular value for the exclusion of LVSD as indicated by the high NPV. We decided not to further reduce the LVSD probability threshold in order to achieve an even higher sensitivity. In a primary care setting, it has already been shown that applying a model for LVSD detection with a comparable sensitivity to an unselected patient cohort led to a significantly higher rate of diagnosing heart failure with reduced ejection fraction than the standard-of-care.² A comparable sensitivity can therefore be considered effective with regard to the clinical applicability of the model. Moreover, lowering the cut-off in order to improve sensitivity would also increase the number of false positives and therefore the number of unnecessary tests as a consequence of the model's output. In this regard, it is important to mention that LVSD prevalence in the population to be tested is of the greatest importance when determining the optimal individual cutoff. Aside from increasing the diagnostic yield, the feasibility of managing cases with potentially identified LVSD in clinical practice as well as the





Figure 4 AUROC as a function of LVEF cut-off for the definition of LVSD.

socioeconomic efficiency must also be included in the discussion. Optimizing models' accuracy would obviously be the best way to reduce the number of false results. Our attempt to improve the model's discriminatory performance by ECG pre-processing was not successful. A pre-selection of clinical high-risk populations and the integration of available basic clinical information as well as NT-proBNP into combined models may be helpful.^{24,25} Furthermore, confirming the reliability of a model by an external validation in unrelated patient cohorts is indispensable in advance of a broad implementation.

There are other methods of advanced ECG analysis for the detection of LVSD that are not based on Al-based pattern recognition, some of which having been reported to have very good discriminatory power.^{28–31} The comparatively high effort required for data analysis could be on possible reason why such algorithms have not yet been widely used. On the other hand, pathophysiological comprehensibility and transparency are positive aspects. In contrast, the Blackbox character of the above-mentioned Al-based pattern recognition models has to be considered a major limitation of them. However, recently there have been attempts to solve the problem of lacking explainability.^{32,33}

Subgroup	Total	Low EF		AUROC (95% C
Age				
20–59 years	12 527	1 477	+	0.917 (0.910–0.
60–69 years	9 780	1 661		0.885 (0.876–0.
70–79 years	11 258	1 932	-	0.863 (0.854–0.
≥ 80 years	8 726	1 513		0.825 (0.814–0.
Sex				
Male	26 106	4 918	+	0.877 (0.873–0.
Female	16 185	1 665	-	0.869 (0.860-0.
Heart rate				
< 60 bpm	6 547	484	- <u>i</u>	0.885 (0.871-0.
60–79 bpm	20 588	2 358		0.883 (0.877-0.
80–99 bpm	10 038	2 025		0.868 (0.859-0.
≥ 100 bpm	5 090	1 710	-	0.810 (0.797–0.
QRS interval				
< 120 ms	30 481	2 957	•	0.885 (0.879–0.
≥ 120 ms	11 782	3 620	-	0.803 (0.794-0.
PQ interval				
100–200 ms	28 635	3 780	+	0.893 (0.888-0.
> 200 ms	7 963	1 514	-	0.863 (0.853-0.
Rhythm				
Sinus rhythm	21 938	3 286	-	0.885 (0.879–0.
Atrial fibrillation/flutter	3 276	649		0.815 (0.798-0.
Other	17 077	2 648	+	0.881 (0.874-0.
Pacemaker				
No	41 704	6 420	÷	0.879 (0.875–0.
Yes	587	163	I	0.798 (0.758–0.
ECG recording				,
All ECGs	42 291	6 583	+	0.878 (0.873–0.
First ECG	31 944	4 780		0.884 (0.879–0.
Last ECG	31 944	4 780	•	0.882 (0.878–0.
Time between ECG and echocardic	ography		1	v
Same day	29 761	4 137	+	0.881 (0.876–0.
1–3 davs	9 814	1 928	-	0.862 (0.853-0.
4–7 davs	2 716	518		0.884 (0.870-0.
All ECGs	42 291	6 583	+	0.878 (0.873-0.

Figure 5 Forrest plot of AUROCs with Cls stratified for baseline and ECG characteristics.

Direct comparisons between the two groups of models for the ECG-based detection of LVSD are currently lacking.

When stratifying results for age, sex, and several ECG parameters, we confirmed the findings provided by Yagi et al.¹⁴ with lower AUROCs in octogenarians, prevalent atrial fibrillation and ECGs with wide QRS complexes. This is in line with findings from other working groups and models.^{26,34,35} Since the clinical diagnoses of atrial fibrillation or existing (left) bundle branch block are likely to be considered as potential indicators of structural heart disease by clinicians, an echo-cardiography will be performed in most cases either way. Therefore, the inferior discriminatory power of AI models related to these ECG patterns should not be relevant in the context of population screening for LVSD. Rather, it can and should be discussed whether separate algorithms should be developed for patients with and without such obvious electrical abnormalities like complete bundle branch block or atrial fibrillation in order to further improve the models' respective

predictions. The fact that AUROCs were lower when definitions of LVSD also included cases with only mildly reduced LVEF has also been shown previously.^{23,36} Of note, we were able to show that patients with a preserved LVEF and a false positive AI-based result for LVSD detection from baseline ECG were at an increased risk for the development of an impaired LVEF during FU compared to patients with a low AI model-computed LVSD probability. Similar observations were made for the model that was presented by colleagues from the Mayo clinic and another recently presented model from Taiwan.^{6,25,37} Of course, all models were not developed for the future detection of LVSD in patients without impaired LVEF at baseline, which might explain the inferior AUROC for this outcome as shown in our analysis. However, even though the performance metrics were lower when compared to corresponding results for LVSD detection at baseline, the huge difference with regard to incidences of LVSD over time has to be considered clinically meaningful. Moreover, after integrating



Figure 6 Incidence of LVSD during FU stratified for model output at baseline in patients with initial LVEF \geq 50%.

Table 1Univariable and multivariable analyses fordevelopment of LVSD during FU in patients with normalLVEF at baseline

Variable	Univariable a	nalysis	Multivariable analysis		
	OR (95% CI)	P value	OR (95% CI)	P value	
LVEF at baseline	0.90 (0.87–0.93)	<0.001	0.92 (0.89–0.96)	<0.001	
Age	1.75 (1.23–2.55)	0.003	1.74 (1.19–2.59)	0.005	
Female sex	0.56 (0.38–0.82)	0.004	0.56 (0.37–0.83)	0.005	
Log model	1.48 (1.34–1.63)	<0.001	1.37 (1.24–1.52)	<0.001	
output score					

Odds ratios are given for female compared to male sex, per 25-year increase of age, per per cent increase in LVEF, and per 1 point change of the logarithmic prediction score. CI, confidence interval; FU, follow-up; LVEF, left ventricular ejection fraction; LVSD, left ventricular systolic dysfunction; OR, odds ratio.

the model's output (probability of LVSD) at baseline with age, sex, and baseline ejection fraction into a multivariable model, the AUROC improved to 0.75, which could definitely serve as a starting point for an intensified clinical FU. In this light, Chen *et al.*¹² reported increased major adverse cardiovascular event rates in patients with a high predicted probability of LVSD despite preserved LVEF at baseline. Further research is needed to assess the additional value of AI-based ECG analysis for the screening of LVSD at baseline and during FU from a clinical and socioeconomical perspective.

Limitations

There are several limitations related to this study. First of all, the technical background of ECG recording as well as the quality of ECG data may be different from the original model development study, which could have influenced results. Due to the sample size of our cohort, a manual quality check of included ECGs was not possible and quality assurance relied on an automated software algorithm for the detection of ECGs with unacceptable quality. Moreover, there was no information on the clinical setting of ECG assessment from the development study by Yagi and colleagues. However, proving the reliability and reproducibility of the model's performance in different patient populations and across different technical requirements is a major goal of external validation. The cut-off value for model output probabilities that was used within the data analyses presented in the original study was not made available by Yagi and colleagues. Therefore, we had to compute an individual cut-off probability with optimized model performance metrics based on our validation dataset. This hinders direct comparability of all described performance measures and is a major limitation. To enable an external validation and further application of an Al-based prediction model, it is of outmost importance to publish the model's architecture including used weights and cut-off values for model's output.

Due to data availability, we were not able to validate the model with prospectively collected data to further add information on reliability and usability as a prediction tool for clinical practice. Furthermore, electronic data on NT-proBNP were available only for a minority of patients from our ECG database, which is why we were not able to compare predictive performance of the examined AI model and NT-proBNP for the detection of LVSD. With regard to the stratification of performance metrics according to different ECG patterns, we relied on the automatic software analysis of ECGs for the detection of atrial fibrillation, wide QRS complexes, and other variables. This carries an unquantifiable risk of misinterpretation, but was unavoidable due to the size of the database. Second, the mode of LVEF assessment from the original study was not described in further detail, which hinders a comparison with our data. Moreover, there may be inconsistencies within our EMR database with respect to LVEF values as they were collected retrospectively and not stored for research purposes. There was no re-evaluation of echocardiography findings. Lastly, there was no regular and scheduled FU for all patients including repeated echocardiography. This is central to the interpretation of LVSD prediction during FU in patients without LVEF impairment at baseline and may have influenced results. A prospective evaluation with planned FU imaging is required to generate more valid data in this regard.

Conclusion

With this study, we provide the external validation of an existing Al-based ECG-analysing model for the detection of LVSD with excellent and robust performance metrics. Several ECG patterns that influenced the model's discrimination were identified. Moreover, patients with preserved LVEF but model-generated increased probability of LVSD at baseline were shown to be at higher risk for a future deterioration of LVSD, which deserves a further evaluation in prospective trials.

Supplementary material

Supplementary material is available at European Heart Journal – Digital Health.

Acknowledgements

The authors thank Dr Ryuichiro Yagi, Prof. Calum A. MacRae, Prof. Rahul C. Deo (all: One Brave Idea & Division of Cardiovascular Medicine, Brigham and Women's Hospital, Boston, MA, USA; Harvard Medical School, Boston, MA, USA), Dr Shinichi Goto (One Brave Idea & Division of Cardiovascular Medicine, Brigham and Women's Hospital, Boston, MA, USA; Harvard Medical School, Boston, MA, USA; Department of Cardiology, Keio University School of Medicine, Shinjuku, Tokyo, Japan) and Dr Yoshinori Katsumata (Keio University School of Medicine, Shinjuku, Tokyo, Japan) for their excellent and innovative work as well as for the possibility of use and apply the algorithm they developed.

Funding

There has been no funding related to the work presented in this manuscript.

Conflict of interest: none declared.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Ethics approval and patients' individual consent

The study complied with the principles outlined in the Declaration of Helsinki and was approved by the responsible local ethics committee of the University of Leipzig (511/19-ek). Due to the retrospective character of the analysis of large, pseudonymized data, no individual informed consent has been obtained.

Permissions information

The authors do hereby declare that all illustrations and figures in the manuscript are entirely original and do not require reprint permission.

References

- Escobar C, Palacios B, Varela L, Gutierrez M, Duong M, Chen H, et al. Healthcare resource utilization and costs among patients with heart failure with preserved, mildly reduced, and reduced ejection fraction in Spain. BMC Health Serv Res 2022;22:1241.
- Odegaard KM, Hallen J, Lirhus SS, Melberg HO, Halvorsen S. Incidence, prevalence, and mortality of heart failure: a nationwide registry study from 2013 to 2016. ESC Heart Fail 2020;7:1917–1926.
- McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Bohm M, et al. 2021 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J 2021;42:3599–3726.
- Averina M, Stylidis M, Brox J, Schirmer H. NT-ProBNP and high-sensitivity troponin T as screening tests for subclinical chronic heart failure in a general population. ESC Heart Fail 2022;9:1954–1962.
- Adedinsewo D, Carter RE, Attia Z, Johnson P, Kashou AH, Dugan JL, et al. Artificial intelligence-enabled ECG algorithm to identify patients with left ventricular systolic dysfunction presenting to the emergency department with dyspnea. *Circ Arrhythm Electrophysiol* 2020;**13**:e008437.
- Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019;25:70–74.
- Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol* 2021;**18**: 465–478.
- Somani S, Russak AJ, Richter F, Zhao S, Vaid A, Chaudhry F, et al. Deep learning and the electrocardiogram: review of the current state-of-the-art. Europace 2021;23:1179–1191.
- Sun JY, Qiu Y, Guo HC, Hua Y, Shao B, Qiao YC, et al. A method to screen left ventricular dysfunction through ECG based on convolutional neural network. J Cardiovasc Electrophysiol 2021;32:1095–1102.
- Attia ZI, Harmon DM, Behr ER, Friedman PA. Application of artificial intelligence to the electrocardiogram. *Eur Heart J* 2021;42:4717–4730.
- Bjerken LV, Ronborg SN, Jensen MT, Orting SN, Nielsen OW. Artificial intelligence enabled ECG screening for left ventricular systolic dysfunction: a systematic review. *Heart Fail Rev* 2023;28:419–430.
- Chen HY, Lin CS, Fang WH, Lou YS, Cheng CC, Lee CC, et al. Artificial intelligenceenabled electrocardiography predicts left ventricular dysfunction and future cardiovascular outcomes: a retrospective analysis. J Pers Med 2022;12:455.
- Chung CT, Lee S, King E, Liu T, Armoundas AA, Bazoukis G, et al. Clinical significance, challenges and limitations in using artificial intelligence for electrocardiography-based diagnosis. Int J Arrhythmia 2022;23:24.

- 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.
- Christopoulos G, Graff-Radford J, Lopez CL, Yao X, Attia Zl, Rabinstein AA, et al. Artificial intelligence-electrocardiography to predict incident atrial fibrillation: a population-based study. Circ Arrhythm Electrophysiol 2020;13:e009355.
- Noseworthy PA, Attia ZI, Brewer LC, Hayes SN, Yao X, Kapa S, et al. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. *Circ Arrhythm Electrophysiol* 2020;**13**:e007988.
- Elias P, Poterucha TJ, Rajaram V, Moller LM, Rodriguez V, Bhave S, et al. Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease. J Am Coll Cardiol 2022;80:613–626.
- Thiele C, Hirschfeld G. cutpointr: improved estimation and validation of optimal cutpoints in R. J Stat Softw 2021;98:27.
- Attia IZ, Tseng AS, Benavente ED, Medina-Inojosa JR, Clark TG, Malyutina S, et al. External validation of a deep learning electrocardiogram algorithm to detect ventricular dysfunction. Int J Cardiol 2021;329:130–135.
- Katsushika S, Kodera S, Nakamoto M, Ninomiya K, Inoue S, Sawano S, et al. The effectiveness of a deep learning model to detect left ventricular systolic dysfunction from electrocardiograms. Int Heart J 2021;62:1332–1341.
- Cho J, Lee B, Kwon JM, Lee Y, Park H, Oh BH, et al. Artificial intelligence algorithm for screening heart failure with reduced ejection fraction using electrocardiography. ASAIO J 2021;67:314–321.
- 22. Kwon JM, Kim KH, Jeon KH, Kim HM, Kim MJ, Lim SM, et al. Development and validation of deep-learning algorithm for electrocardiography-based heart failure identification. *Korean Circ J* 2019;**49**:629–639.
- Vaid A, Johnson KW, Badgeley MA, Somani SS, Bicak M, Landi I, et al. Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. JACC Cardiovasc Imaging 2021;15:395–410.
- Attia ZI, Kapa S, Yao X, Lopez-Jimenez F, Mohan TL, Pellikka PA, et al. Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction. J Cardiovasc Electrophysiol 2019;30:668–674.
- Kashou AH, Medina-Inojosa JR, Noseworthy PA, Rodeheffer RJ, Lopez-Jimenez F, Attia IZ, et al. Artificial intelligence-augmented electrocardiogram detection of left ventricular systolic dysfunction in the general population. Mayo Clin Proc 2021;96:2576–2586.
- Perez-Downes J, Fitzgerald P, Adedinsewo D, Carter RE, Noseworthy PA, Kusumoto F. Impact of ECG characteristics on the performance of an artificial intelligence enabled ECG for predicting left ventricular dysfunction. *Circ Arrhythm Electrophysiol* 2021;14:e009871.
- Yao X, Rushlow DR, Inselman JW, McCoy RG, Thacher TD, Behnken EM, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med 2021;27:815–819.
- Johnson K, Neilson S, To A, Amir N, Cave A, Scott T, et al. Advanced electrocardiography identifies left ventricular systolic dysfunction in non-ischemic cardiomyopathy and tracks serial change over time. J Cardiovasc Dev Dis 2015;2:93–107.
- Rautaharju PM, Kooperberg C, Larson JC, LaCroix A. Electrocardiographic predictors of incident congestive heart failure and all-cause mortality in postmenopausal women: the Women's Health Initiative. *Circulation* 2006;**113**:481–489.
- Rautaharju PM, Zhang ZM, Haisty WK Jr, Prineas RJ, Kucharska-Newton AM, Rosamond WD, et al. Electrocardiographic predictors of incident heart failure in men and women free from manifest cardiovascular disease (from the Atherosclerosis Risk in Communities [ARIC] study). Am J Cardiol 2013;112:843–849.
- Schlegel TT, Kulecz WB, Feiveson AH, Greco EC, DePalma JL, Starc V, et al. Accuracy of advanced versus strictly conventional 12-lead ECG for detection and screening of coronary artery disease, left ventricular hypertrophy and left ventricular systolic dysfunction. BMC Cardiovasc Disord 2010;10:28.
- Katsushika S, Kodera S, Sawano S, Shinohara H, Setoguchi N, Tanabe K, et al. An explainable artificial intelligence-enabled electrocardiogram analysis model for the classification of reduced left ventricular function. Eur Heart J Digit Health 2023;4:254–264.
- 33. van de Leur RR, Bos MN, Taha K, Sammani A, Yeung MW, van Duijvenboden S, et al. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. Eur Heart J Digit Health 2022;3:390–404.
- 34. Harmon DM, Carter RE, Cohen-Shelly M, Svatikova A, Adedinsewo DA, Noseworthy PA, 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.
- Jentzer JC, Kashou AH, Attia ZI, Lopez-Jimenez F, Kapa S, Friedman PA, 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.
- Golany T, Radinsky K, Kofman N, Litovchik I, Young R, Monayer A, 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.
- Huang YC, Hsu YC, Liu ZY, Lin CH, Tsai R, Chen JS, et al. Artificial intelligence-enabled electrocardiographic screening for left ventricular systolic dysfunction and mortality risk prediction. Front Cardiovasc Med 2023;10:1070641.