HHS Public Access

Author manuscript

Eur Heart J Digit Health. Author manuscript; available in PMC 2022 October 13.

Published in final edited form as:

Eur Heart J Digit Health. 2022 June ; 3(2): 238–244. doi:10.1093/ehjdh/ztac028.

Real-world performance, long-term efficacy, and absence of bias in the artificial intelligence enhanced electrocardiogram to detect left ventricular systolic dysfunction

David M. Harmon[1], **Rickey E. Carter**[2], **Michal Cohen-Shelly**[3], **Anna Svatikova**[4], **Demilade A. Adedinsewo**[5], **Peter A. Noseworthy**[3], **Suraj Kapa**[3], **Francisco Lopez-Jimenez**[3], **Paul A. Friedman**[3], **Zachi I. Attia**[3,*]

[1]Department of Internal Medicine, Mayo Clinic School of Graduate Medical Education, Rochester, MN;

[2]Department of Quantitative Health Sciences, Mayo Clinic College of Medicine, Jacksonville, FL;

[3]Department of Cardiovascular Medicine, Mayo Clinic College of Medicine, Rochester, MN;

[4]Department of Cardiovascular Medicine, Mayo Clinic College of Medicine, Scottsdale, AZ;

[5]Department of Cardiovascular Medicine, Mayo Clinic College of Medicine, Jacksonville, FL

Abstract

Aims—Some artificial intelligence models applied in medical practice require ongoing retraining, introduce unintended racial bias, or have variable performance among different subgroups of patients. We assessed the real-world performance of the artificial intelligence-enhanced electrocardiogram to detect left ventricular systolic dysfunction with respect to multiple patient and electrocardiogram variables to determine the algorithm's long-term efficacy and potential bias in the absence of retraining.

Methods and results—Electrocardiograms acquired in 2019 at Mayo Clinic in Minnesota, Arizona, and Florida with an echocardiogram performed within 14 days were analyzed (n = 44 986 unique patients). The area under the curve (AUC) was calculated to evaluate performance of the algorithm among age groups, racial and ethnic groups, patient encounter location, electrocardiogram features, and over time. The artificial intelligence-enhanced electrocardiogram to detect left ventricular systolic dysfunction had an AUC of 0.903 for the total cohort. Time series analysis of the model validated its temporal stability. Areas under the curve were similar for all racial and ethnic groups (0.90–0.92) with minimal performance difference between sexes. Patients with a 'normal sinus rhythm' electrocardiogram (n = 37 047) exhibited an AUC of 0.91. All other electrocardiogram features had areas under the curve between 0.79 and 0.91, with the lowest performance occurring in the left bundle branch block group (0.79).

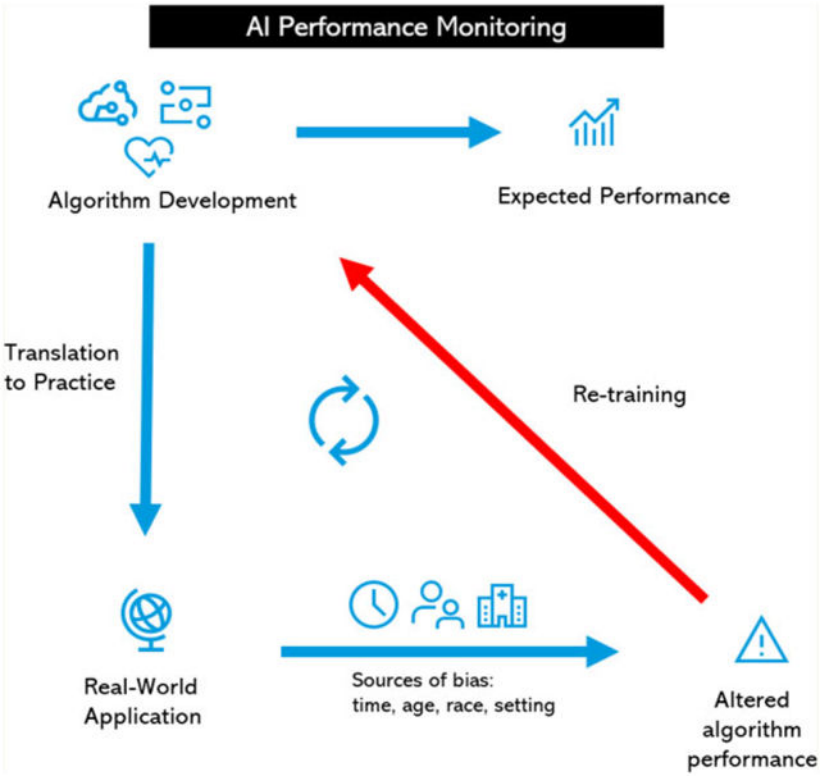
*Corresponding author. Fax: 507-255-2550, attia.itzhak@mayo.edu.

Supplementary material

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

Conclusion—The artificial intelligence-enhanced electrocardiogram to detect left ventricular systolic dysfunction is stable over time in the absence of retraining and robust with respect to multiple variables including time, patient race, and electrocardiogram features.

Graphical Absract



AI Performance Monitoring.

Keywords

Artificial intelligence; ECG; Heart failure; Arrhythmia; Digital medicine; Deep learning

Introduction

Left ventricular systolic dysfunction (LVSD) is a common pathology associated with significant morbidity and mortality.[1–3] Although no routine screening strategies are currently recommended, robust efforts have been made to identify patients with LVSD.[4–6] Early identification may permit treatment with evidence-based therapies to mitigate morbidity and mortality.[1]

The electrocardiogram (ECG) is a rapid, cost-effective, point-of-care test that is available in most clinical settings, both inpatient and outpatient. We have previously reported the development and clinical application of a deep learning network to detect LVSD using an artificial intelligence-enhanced ECG (AI ECG) in patients in primary care clinics and in the emergency department.[7–10] The AI ECG remains promising as a non-invasive screening tool

for patients with previously undiagnosed LVSD, including those with suspected myocardial dysfunction secondary to SARS-CoV-2 infection,[11] for which it has received US Food and Drug Administration Emergency Use Authorization.

It is well known that a trained network is susceptible to data set shift errors when employed in a setting that differs from its original training environment. In some cases, AI models have to re-learn or recalibrate to better fit an environment or setting.[12] Although our group has performed both internal and external validation of the AI ECG algorithm to detect LVSD with promising results,[13,14] its stability with respect to potentially impactful clinical factors (i.e. patient age, ECG recording location, underlying arrhythmia, time since network training) remains undetermined. To test the hypothesis that the AI ECG network for LVSD is stable over time and across populations due to its broad training set, we assessed its accuracy with respect to time since development, patient age, patient race and ethnicity, ECG rhythm and morphology, and ECG recording location.

Methods

Study population

All ECGs acquired between 1 January and 31 December 2019 (the year following algorithm development) performed at Mayo Clinic in Rochester, Minnesota, Scottsdale, Arizona, Jacksonville, Florida, and Mayo Clinic Health System locations were evaluated (n = 406 916 from 230 661 patients). Eligible ECGs were digital, standard 10 s 12-lead ECGs acquired in the supine position during the study timeframe. ECGs from patients who underwent transthoracic echocardiography within 2 weeks of an eligible ECG were included in this analysis. In the case of multiple echocardiograms for a single patient, the earliest echocardiogram was selected as the index echocardiogram. In the case of multiple ECGs in the 14-day window from echo, the ECG temporally closest to the first echo was selected for analysis (n = 44 986 patients with valid ECG-echo pairs). None of the patients in this analysis participated in the derivation of algorithm, and all ECGs in the analysis were acquired after the model was fully developed.

Patient data

The institutional review boards of the Mayo Clinic Foundation approved of this study and protocol. Following institutional review board approval, patient age and self-reported race and ethnicity were collected from the Mayo Clinic unified data platform. The original protocol was modified to include patients <18 years of age, and the institutional review board approved this change. The ECG date, acquisition location, and rhythm were extracted from the MUSE system (GE Healthcare, Marquette, WI). For all ECGs, the final rhythm and other ECG findings were adjudicated by a technologist under cardiology supervision.

All echocardiograms had one or more ejection fraction (EF) measurement performed by a cardiologist. For studies with more than one left ventricular ejection fraction measurement, we used a heuristic technique to select the most accurate measurement.[7] The preferred measurement used for analysis was (from most accurate to least accurate): 3D echocardiography, biplane imaging using the Simpson method, 2D methods, M-mode

measurements, and, in the absence of any of the preceding, the reported visually estimated EF.

Artificial intelligence model

The artificial intelligence model used in this study has been previously described.[7] Briefly, the model uses a convolutional neural network that analyzes a matrix with 10 s, 12-lead ECG data resampled to 500 Hz (5000 × 12 values per ECG). Each matrix row contains the raw amplitude for each of the 12 leads for that timestamp. The model uses seven convolutional blocks, each with a convolutional layer, batch normalization, a 'Relu' activation function and a max-pooling layer, followed by two fully connected blocks.[7] The model was developed using 44 959 unique patients and was tested on 52 870 patients not used to develop the model. In the original testing cohort, the model was able to detect an EF ≤ 35% with an area under the curve (AUC) of 0.93 and an EF ≤ 40% with an AUC of 0.91.

Although the model was developed and tested prior to the start date of the present study, to avoid any data leakage, patients used to develop the model were excluded from the current analysis, as we assume that the model might have an unfair advantage when testing the same patients again due to the biometric information in the ECG.

Main outcome

While the original publication focused on the detection of EF ≤ 35%, in this work we adjusted the threshold for the definition of low EF to ≤ 40%, in accordance with the Universal Definition and Classification of Heart Failure, European Society of Cardiology, and American College of Cardiology professional society guidelines.[1,15,16] The AUC for the detection of EF ≤ 40%, was evaluated with respect to multiple factors: time (in 1-month blocks, for each month of 2019), patient age (in 10-year intervals), race and ethnicity, ECG recording location and situation (inpatient, outpatient or emergency department), echocardiogram measurement type used to assess EF, and ECG features at the time of screening.

The focus of the statistical analysis was providing descriptive data on the performance of the algorithm over the range of patient profiles identified above. No formal hypothesis testing was configured to test for differences from the original area under the receiver operating characteristic curve. Instead, differences in the point estimates of AUC between the original sample and the new validation cases greater than 3 percentage points were considered because the original study yielded confidence intervals within the precision of 1 percentage point.[7] An exploratory analysis was conducted by calculating month-by-month estimates of the AUC over the 12-month study period. Data were also presented in forest plots that provided an estimate of sensitivity and specificity. To arrive at binary test predictions, the previously selected threshold of 0.256 was utilized (i.e. a new threshold optimized for EF ≤ 40% was not considered in the analysis). Data analysis was conducted using Python version 3.76 and R version 4.0.3.

Results

Background characteristics

The overall cohort had a mean age of 64.0 ± 17.6 years, 43% were female, and 9.6% of the population had an ejection fraction (EF) of ≤ 40%. The mean absolute time between the screening ECG and the echocardiogram was 1.5 ± 2.2 days. The majority of patients were white (89.2%), and patients with an EF ≤ 40% were significantly older and more comorbid than patients with an EF >40%. Details of age, race, and comorbidities by EF value are described in Table 1. Expanded race and ethnicity characteristics are described in Supplementary material online, Table S1.

Temporal evaluation

Evaluation of the AUC by month is shown in Figure 1. The per-month AUC remains greater than 0.89 throughout the year of 2019 without significant variation. There is moderate variability of the confidence interval range on a month-to-month basis; however this does not significantly impact overall accuracy of the algorithm over time. This visual finding was supported with the time series model. In particular, the estimated slope of the performance over the 12-month period was 0.0002 consistent with temporal stability of the model.

Age group and sex evaluation

Evaluation of AUC with respect to age revealed robust test performance (AUC >0.87) for the age groups 0–10 years old and 20–80 years old (Figure 2A). The AUC was lower for patients with 10–20 years of age and greater than 80 years (AUC 0.831–0.843); however, the sample size of the number of patients aged under 20 years of age was small (n = 1183). The AUC was minimally different between male and female patients (Figure 2B).

Electrocardiogram accuracy with respect to race and ethnicity

To validate that the model is not impacted by patient race, we used self-reported race and ethnicity to determine whether significant differences in AUC among the groups exist. In this validation cohort, 43 302 patients self-reported race and ethnicity. The model had similar AUC among all races and ethnicities tested (AUC: 0.90–0.93) with slightly higher AUC for black and Hispanic patients, as noted in Figures 2C, 2D, and Supplementary material online, Figure S1. Patients with unknown racial demographics had slightly lower AUC (0.899; see Supplementary material online, Figure S1).

Electrocardiogram accuracy with respect to ejection fraction measurement type

For each echocardiogram the most accurate EF measurement available was used (order of accuracy from highest to lowest: 3D, Simpson method, 2D method, M-mode, and visual estimation of the EF). Electrocardiogram evaluation of EF exhibited a higher AUC when using 3D EF estimates from the echocardiogram (AUC 0.951) and lower AUC when EF was obtained by visual estimation (0.743) when compared with 2D-based modalities (AUCs ranged from: 0.89–0.94 with overlapping confidence internals, Figure 2E).

Electrocardiogram recoding location

Evaluation of the AI algorithm performance with respect to collection site and location of appointment is described in detail in Figure 2F and 2G. With respect to appointment location, outpatients had the highest AUC (0.925), followed by the emergency department (0.897) and inpatient locations (0.866) (Table 2F). In Figure 2G, minimal AUC differences among the different Mayo Clinic locations were noted (AUC 0.903–0.909).

Area under the curve by electrocardiogram features

AUC characteristics stratified by ECG features are depicted in Figure 3. Electrocardiogram features/findings were evaluated on an individual basis as some ECG features are mutually exclusive (NSR and atrial flutter for example). The overall evaluation for all analyzed ECGs had an AUC of 0.903 (2948 of 44919 with LVSD), similar to the AUC for detection of low EF in our original work (AUC = 0.91 for EF ≤ 40%).[7] Patients with an ECG rhythm of 'normal sinus rhythm' exhibited an AUC of 0.91 for detecting EF ≤ 40%. All other ECG rhythms or ECG diagnoses had AUCs between 0.791 (for left bundle branch block) and 0.91, as shown in Figure 3.

Discussion

We found that the AI ECG model for LVSD was robust without retraining even for the less stringent definition of LVSD of ≤ 40% EF. Specifically, it was stable over time, and functionally invariant with regards to patient sex and race. There was some minor variability in performance based on age, ECG recording location, and ECG features that may warrant further consideration.

Neural networks are susceptible to data shift errors when networks are applied to populations that differ from those initially used for training.[12] Although the LVSD AI ECG neural network was trained on a large and broad population, and previous retrospective analysis had shown robustness with regards to patient age and ethnicity,[17] this is the first real-world prospective use study to show temporal stability, and robustness over a wide range of geographical usage (with inherently different populations), clinical environments (outpatient, inpatient, emergency department), and ECG features. Through each of these evaluations we identified only minor variations in algorithm accuracy across each clinically important variable. This non-invasive point-of-care screening tool exhibits significant dataset stability on multiple fronts.

The minimal variation of AUC over time in 2019 suggests that the algorithm does not require recalibration on a monthly or annual basis. These data suggest that the AI ECG for LVSD is minimally susceptible to temporal dataset shifts. However, an underlying assumption is the use of a standard 12-lead ECG acquired using clinically accepted lead positions with the patient supine. The use of ECG signals from mobile or other form factors requires additional validation, as variations in cardiac position, adrenergic tone, resting and active heart rates, and ECG morphology could impact algorithm function. Although this evaluation indicates no anticipated performance degradation over time, re-evaluation in

future years ssmay be also merited in the event of shifting disease trends, population change, or alterations in the ECG system.

The AI ECG is considered a black box, since the specific ECG features used to make EF determination is unknown. Thus, there is a concern for the introduction of implicit bias, especially as it relates to underrepresented groups. Prior to using the model in practice, we demonstrated that in our testing cohort, the LVSD model worked well in all races and ethnicities in the testing set.[17] In this prospective validation, we found that despite the known ECG features differences across races and ethnicities, the LVSD algorithm works equally well in those that were tested. Similarly, we observed robust performance of the algorithm for both male and female patients.

There was a modest drop in test performance with advanced age, with an AUC of 0.87 for patients 80–90, and 0.84 for age above 90 (Figure 2A). Noteworthy is that the most clinically relevant age group for LVSD (20–80 years of age) has the greatest test performance (AUC >0.89). Irrespective of the performance degradation with advanced age, the test showed robust performance characteristics with an AUC >0.8 for all age groups with values remaining in line with other medical tests (typically AUC > 0.75). Although we observed an acceptable AUC in patients 0–20 years, the number of patients with disease (n = 13) was quite small in these groups, the sensitivity for patients aged <20 years was low (37.5–60.0%), and this experimental data represented a first-time application of the AI ECG for LVSD algorithm, in training or testing, to a paediatric patient population. As a result of these limitations, no conclusions can be drawn regarding algorithm use in these low prevalence, paediatric populations. Further investigation in a larger paediatric population must be completed to clarify appropriate application of this algorithm.

When validating LVSD accuracy by echocardiography, the AUC significantly decreased when the EF was measured by visual estimation (AUC <0.8). All other modes of measuring EF by echocardiography exhibited AUC greater than 0.9 (except 2D; AUC = 0.889), suggesting that the accuracy EF by visual estimation may be limited. This is consistent with previous reports describing the variability of EF measurement via visual estimation.[18] User variability and clinical context (i.e. severely ill in emergency department with point-of-care echocardiogram) likely contribute to the variation in test performance. This raises the possibility that the AI ECG may, at times, be a more powerful predictor of LVSD than echocardiography when ventricular function is visually estimated, particularly in acute environments and clinical situations in which echocardiographic images may be challenging. However, given the limited number of patients in this subgroup (n = 133), it is difficult to draw definite conclusions without further investigation.

Location analysis reveals that the AUC was higher in the outpatient setting than in the emergency department or inpatient setting. This heightened test performance may reflect fewer confounding factors present in the clinic setting. A number of hypotheses may explain this phenomenon: (i) ambulatory patients may be less likely to have multiple active medical issues or an acute exacerbation of another disease, either of which might impact the ECG; (ii) ambulatory patients typically do not have external factors impacting clinical haemodynamics (i.e. intubation, positive pressure ventilation, use of intravenous vasoactive

medications); (iii) there is likely less variation between ECG machines used in the clinic (i.e. limited number of machines at an outpatient location rather than next available ECG tech within the hospital/unit).

Finally, we hypothesized that variations in ECG rhythm and morphological features likely used by the model may impact its performance. In this analysis we sought to determine whether the presence of a paced rhythm or an intrinsic left bundle branch block (LBBB) are significantly used by the model to identify LVSD (in which case, the AUC would approximate 0.5 in the context of LBBB, as not all individuals with LBBB actually have LVSD). However, irrespective of the absence or presence of specific ECG features, the model continued to perform with an acceptable AUC >0.8. This indicates that the model continues to effectively identify left ventricular dysfunction despite the presence of bundle branch blockade, hypertrophy, or other features listed in Figure 3. Similarly, model performance persisted in the presence of arrhythmias (Figure 3). While patients with ‘completely normal ECGs’ (i.e. absence of any arrhythmia, conduction delay, or other electrophysiologic abnormality) exhibited an AUC of 0.803, the sensitivity was quite low (19.4%) with very high specificity (99.1%) indicating the algorithm may have significant limitation as a screening tool in this subset of patients without adjusting the model threshold to have a higher sensitivity and lower specificity for patients with completely normal ECG.

Limitations

The overall number of underrepresented minorities was small. Patients from across the United States seen at all three main Mayo Clinic campuses and the Midwest health system were included to mitigate this risk. Owing to the study size, although proportionally underrepresented, individuals in each minority group numbered in the hundreds to thousands (see Supplementary material online, Table S1). We acknowledge that while the preliminary performance across race and ethnicity shows promise, further evaluation is required.

Conclusions

The AI ECG is robust screening tool for ventricular dysfunction, with strong performance over time, geography, clinical location, patient age, and ECG features. We similarly observed strong AI ECG performance with respect to variable race/ethnicity while acknowledging the limited diversity of this study population. The change in model performance with method of EF estimation emphasizes the need for objective, reproducible EF calculation methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest:

The AI ECG algorithm to detect left ventricular dysfunction was licensed by Mayo Clinic to Anumana, Eko health. PAF, ZIA, FLJ, REC, SJA and other inventors and advisors to these entities may benefit financially from their commercialization. PAN receives research funding from National Institutes of Health (NIH, including the National Heart, Lung, and Blood Institute [NHLBI, R21AG 62580-1, R01HL 131535-4, R01HL 143070-2] the National Institute on Aging [NIA, R01AG 062436-1]), Agency for Healthcare Research and Quality (AHRQ, R01HS 25402-3), Food and Drug Administration (FDA, FD 06292), and the American Heart Association (18SFRN34230146, AHA). DMH receives support from the NIH StARR Resident Investigator Award (NIH 5R38HL150086-02).

Biography

David M. Harmon is an academic internist at the Mayo Clinic in Rochester, MN, USA. He is a member of the Alpha Omega Alpha Medical Honor Society and is directing research efforts at the Mayo Clinic through a National Institute of Health Stimulating Access to Research in Residency grant. His research is focused on the clinical development and application of artificial intelligence to the electrocardiogram. He aims to broaden the use of artificial intelligence technology, validating its application in wearable and handheld devices which hold implications for both at-home medical care and acute, emergent settings.

Data availability

Our raw data from this study is not available, however the analysis of our results are sharable.

References

1. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE Jr, Drazner MH, Fonarow GC, Geraci SA, Horwich T, Januzzi JL, Johnson MR, Kasper EK, Levy WC, Masoudi FA, McBride PE, McMurray JJV, Mitchell JE, Peterson PN, Riegel B, Sam F, Stevenson LW, Tang WHW, Tsai EJ, Wilkoff BL, American College of Cardiology Foundation, American Heart Association Task Force on Practice Guidelines. 2013 ACCF/AHA Guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 2013;62: e147–e239. [PubMed: 23747642]
2. Wehner GJ, Jing L, Haggerty CM, Suever JD, Leader JB, Hartzel DN, Kirchner HL, Manus JNA, James N, Ayar Z, Gladding P, Good CW, Cleland JGF, Fornwalt BK. Routinely reported ejection fraction and mortality in clinical practice: where does the nadir of risk lie? Eur Heart J 2020;41:1249–1257. [PubMed: 31386109]
3. Pfeffer MA, Braunwald E, Moye LA, Basta L, Brown EJ, Cuddy TE, Davis BR, Geltman EM, Goldman S, Flaker GC, Klein M, Lamas GA, Packer M, Rouleau J, Rouleau JL, Rutherford J, Wertheimer JH, Hawkins CM, The SAVE Investigators. Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction. Results of the survival and ventricular enlargement trial. N Engl J Med 1992;327:669–677. [PubMed: 1386652]
4. Betti I, Castelli G, Barchielli A, Beligni C, Boscherini V, De Luca L, Messeri G, Gheorghiade M, Maisel A, Zuppiroli A, The PROBE-HF study. The role of N-terminal PRO-brain natriuretic peptide and echocardiography for screening asymptomatic left ventricular dysfunction in a population at high risk for heart failure. J Card Fail 2009;15:377–384. [PubMed: 19477397]

5. Redfield MM, Rodeheffer RJ, Jacobsen SJ, Mahoney DW, Bailey KR, Burnett JC Jr. Plasma brain natriuretic peptide to detect preclinical ventricular systolic or diastolic dysfunction: a community-based study. Circulation 2004;109:3176–3181. [PubMed: 15184280]

6. McDonagh TA, McDonald K, Maisel AS. Screening for asymptomatic left ventricular dysfunction using B-type natriuretic Peptide. Congest Heart Fail 2008;14:5–8. [PubMed: 18772638]

7. 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. [PubMed: 30617318]

8. 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. Circ Arrhythm Electrophysiol 2020;13:e008437. [PubMed: 32986471]

9. Yao X, Rushlow DR, Inselman JW, McCoy RG, Thacher TD, Behnken EM, Bernard ME, Rosas SL, Akfaly A, Misra A, Molling PE, Krien JS, Foss RM, Barry BA, Siontis KC, Kapa S, Pellikka PA, Lopez-Jimenez F, Attia ZI, Shah ND, Friedman PA, Noseworthy PA. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med 2021; 27:815–819. [PubMed: 33958795]

10. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol 2021;18: 465–478. [PubMed: 33526938]

11. Attia ZI, Kapa S, Noseworthy PA, Lopez-Jimenez F, Friedman PA. Artificial intelligence ECG to detect left ventricular dysfunction in COVID-19: a case series. Mayo Clin Proc 2020;95:2464–2466. [PubMed: 33153634]

12. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics 2020;21:345–352. [PubMed: 31742354]

13. 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. J Cardiovasc Electrophysiol 2019;30:668–674. [PubMed: 30821035]

14. 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. [PubMed: 33400971]

15. Bozkurt B, Coats AJ, Tsutsui H, Abdelhamid M, Adamopoulos S, Albert N, Anker SD, Atherton J, Böhm M, Butler J, Drazner MH, Felker GM, Filippatos G, Fonarow GC, Fiuzat M, Gomez-Mesa JE, Heidenreich P, Imamura T, Januzzi J, Jankowska EA, Khazanie P, Kinugawa K, Lam CSP, Matsue Y, Metra M, Ohtani T, Francesco Piepoli M, Ponikowski P, Rosano GMC, Sakata Y, SeferoviĆ P, Starling RC, Teerlink JR, Vardeny O, Yamamoto K, Yancy C, Zhang J, Zieroth S. Universal definition and classification of heart failure: a report of the heart failure society of america, heart failure association of the european society of cardiology, japanese heart failure society and writing committee of the universal definition of heart failure. J Card Fail 2021. S1071-9164(21)00050-6.

16. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Böhm M, Burri H, Butler J,Čelutkiene J, Chioncel O, Cleland JGF, Coats AJS, Crespo-Leiro MG, Farmakis D, Gilard M, Heymans S, Hoes AW, Jaarsma T, Jankowska EA, Lainscak M, Lam CSP, Lyon AR, McMurray JJV, Mebazaa A, Mindham R, Muneretto C, Francesco Piepoli M, Price S, Rosano GMC, Ruschitzka F, Kathrine Skibelund A, de Boer RA, Christian Schulze P., Abdelhamid M, Aboyans V, Adamopoulos S, Anker SD, Arbelo E, Asteggiano R, Bauersachs J, Bayes-Genis A, Borger MA, Budts W, Cikes M, Damman K, Delgado V, Dendale P, Dilaveris P, Drexel H, Ezekowitz J, Falk V, Fauchier L, Filippatos G, Fraser A, Frey N, Gale CP, Gustafsson F, Harris J, Iung B, Janssens S, Jessup M, Konradi A, Kotecha D, Lambrinou E, Lancellotti P, Landmesser U, Leclercq C, Lewis BS, Leyva F, Linhart A, Løchen M-L, Lund LH, Mancini D, Masip J, Milicic D, Mueller C, Nef H, Nielsen J-C, Neubeck L, Noutsias M, Petersen SE, Sonia Petronio A, Ponikowski P, Prescott E, Rakisheva A, Richter DJ, Schlyakhto E, Seferovic P, Senni M, Sitges M, Sousa-Uva

M, Tocchetti CG, Touyz RM, Tschoepe C, Waltenberger J, ESC Scientific Document Group. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). With the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail 2022 Jan; 24:4–131. [PubMed: 35083827]

17. 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. Circ Arrhythm Electrophysiol 2020;13:e007988. [PubMed: 32064914]
18. Cole GD, Dhutia NM, Shun-Shin MJ, Willson K, Harrison J, Raphael CE, Zolgharni M, Mayet J, Francis DP. Defining the real-world reproducibility of visual grading of left ventricular function and visual estimation of left ventricular ejection fraction: impact of image quality, experience and accreditation. Int J Cardiovasc Imaging 2015;31: 1303–1314. [PubMed: 26141526]

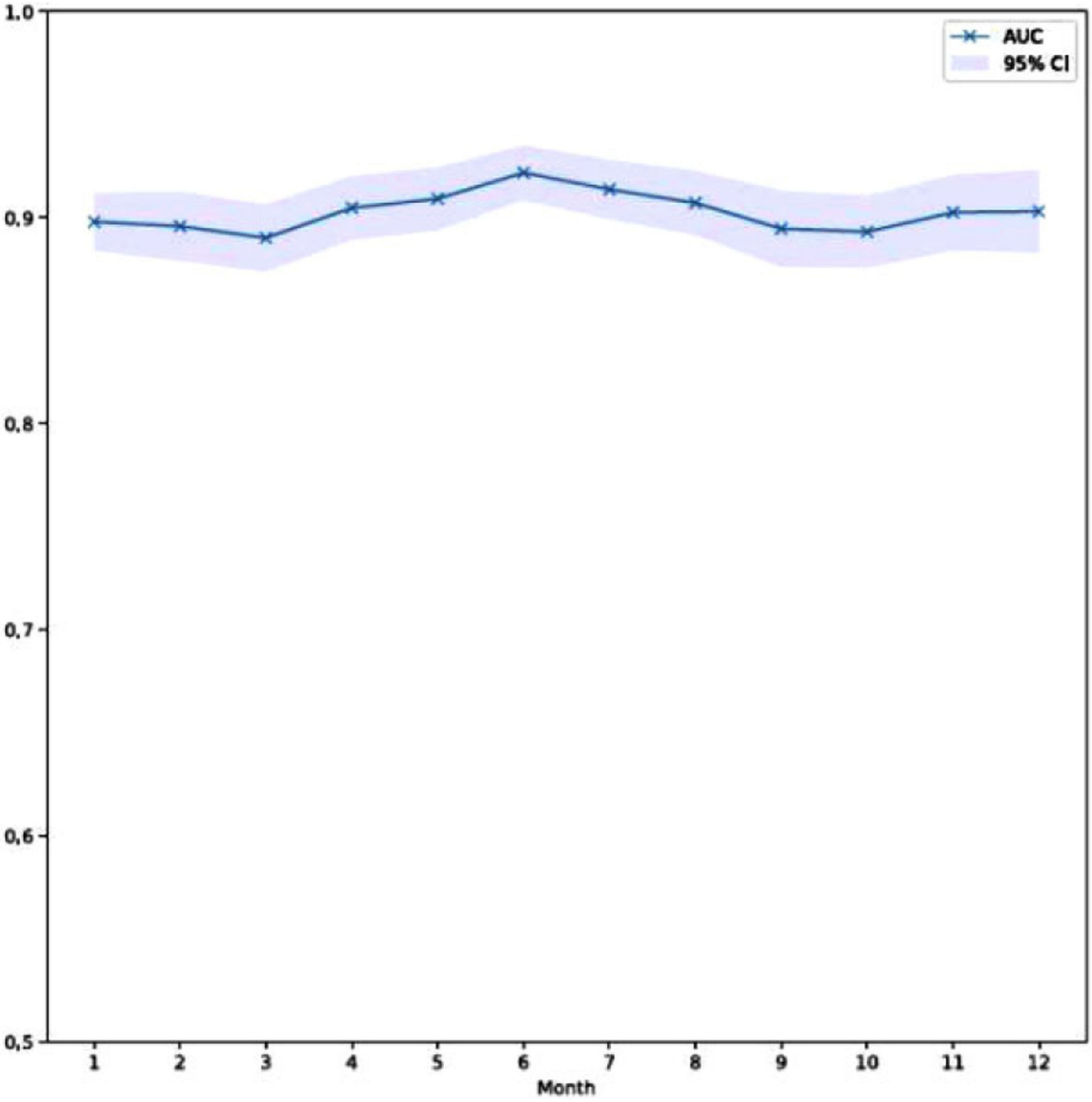
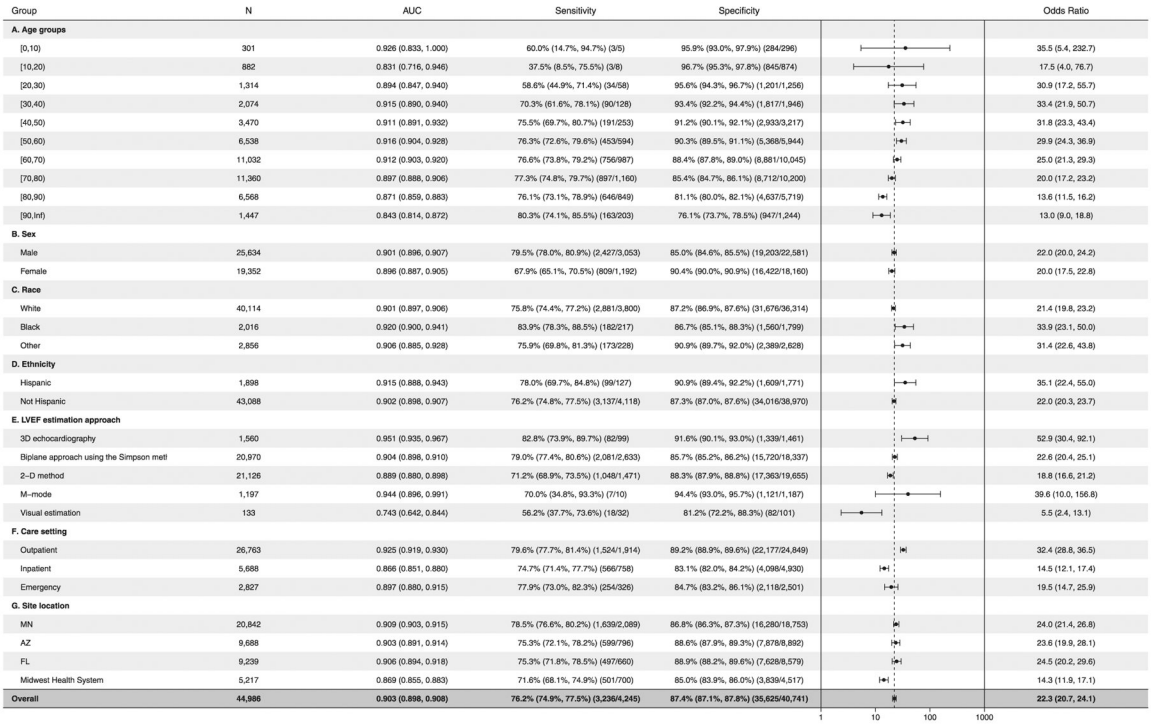


Figure 1.

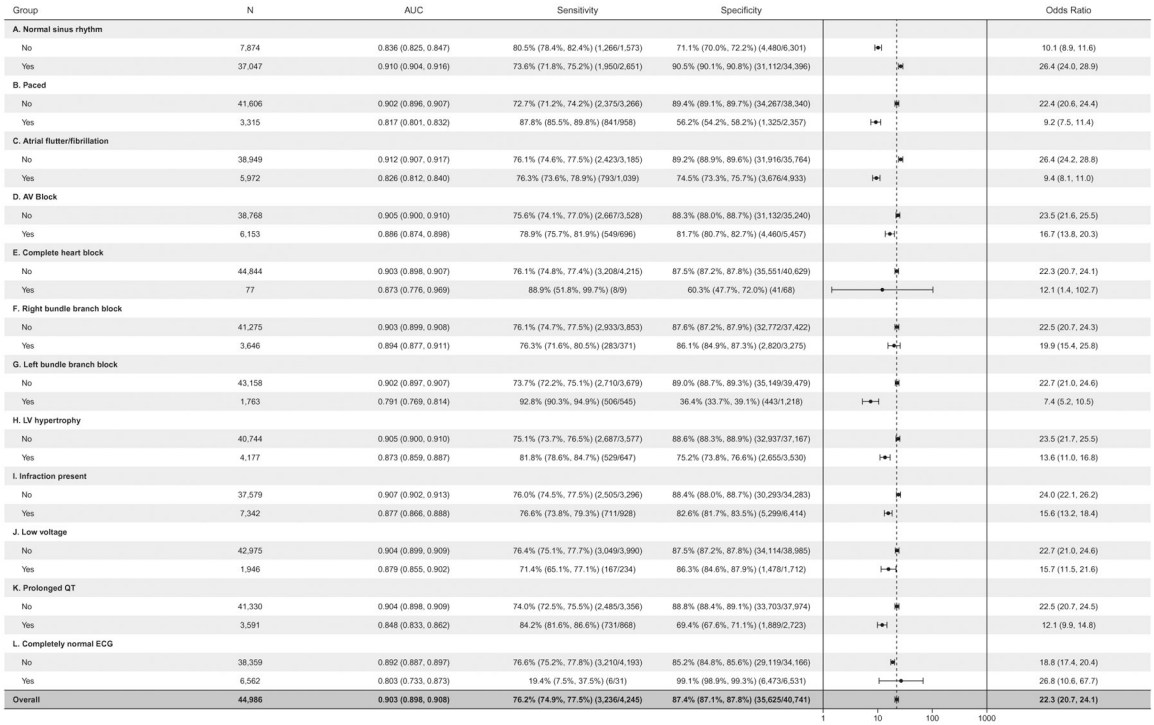
Artificial intelligence-enhanced electrocardiogram for left ventricular systolic dysfunction area under the curve by month in the year of 2019.



Group	N	AUC	Sensitivity	Specificity	Odds Ratio
A. Age groups					
[0,10)	301	0.926 (0.833, 1.000)	60.0% (14.7%, 94.7%) (3/5)	95.9% (93.0%, 97.9%) (284/296)	35.5 (5.4, 232.7)
[10,20)	882	0.831 (0.716, 0.946)	37.5% (8.5%, 75.5%) (3/8)	96.7% (95.3%, 97.8%) (845/874)	17.5 (4.0, 76.7)
[20,30)	1,314	0.894 (0.847, 0.940)	58.6% (44.9%, 71.4%) (34/58)	95.6% (94.3%, 96.7%) (1,201/1,256)	30.9 (17.2, 55.7)
[30,40)	2,074	0.915 (0.890, 0.940)	70.3% (61.6%, 78.1%) (90/128)	93.4% (92.2%, 94.4%) (1,817/1,946)	33.4 (21.9, 50.7)
[40,50)	3,470	0.911 (0.891, 0.932)	75.5% (69.7%, 80.7%) (191/253)	91.2% (90.1%, 92.1%) (2,933/3,217)	31.8 (23.3, 43.4)
[50,60)	6,538	0.916 (0.904, 0.928)	76.3% (72.6%, 79.6%) (453/594)	90.3% (89.5%, 91.1%) (5,368/5,944)	29.9 (24.3, 36.9)
[60,70)	11,032	0.912 (0.903, 0.920)	76.6% (73.8%, 79.2%) (756/987)	88.4% (87.8%, 89.0%) (8,881/10,045)	25.0 (21.3, 29.3)
[70,80)	11,360	0.897 (0.888, 0.906)	77.3% (74.8%, 79.7%) (897/1,160)	85.4% (84.7%, 86.1%) (8,712/10,200)	20.0 (17.2, 23.2)
[80,90)	6,568	0.871 (0.859, 0.883)	76.1% (73.1%, 78.9%) (646/849)	81.1% (80.0%, 82.1%) (4,637/5,719)	13.6 (11.5, 16.2)
[90,Inf)	1,447	0.843 (0.814, 0.872)	80.3% (74.1%, 85.5%) (163/203)	76.1% (73.7%, 78.5%) (947/1,244)	13.0 (9.0, 18.8)
B. Sex					
Male	25,634	0.901 (0.896, 0.907)	79.5% (78.0%, 80.9%) (2,427/3,053)	85.0% (84.6%, 85.5%) (19,203/22,581)	22.0 (20.0, 24.2)
Female	19,352	0.896 (0.887, 0.905)	67.9% (65.1%, 70.5%) (809/1,192)	90.4% (90.0%, 90.9%) (16,422/18,160)	20.0 (17.5, 22.8)
C. Race					
White	40,114	0.901 (0.897, 0.906)	75.8% (74.4%, 77.2%) (2,881/3,800)	87.2% (86.9%, 87.6%) (31,676/36,314)	21.4 (19.8, 23.2)
Black	2,016	0.920 (0.900, 0.941)	83.9% (78.3%, 88.5%) (182/217)	86.7% (85.1%, 88.3%) (1,560/1,799)	33.9 (23.1, 50.0)
Other	2,856	0.906 (0.885, 0.928)	75.9% (69.8%, 81.3%) (173/228)	90.9% (89.7%, 92.0%) (2,389/2,628)	31.4 (22.6, 43.8)
D. Ethnicity					
Hispanic	1,898	0.915 (0.888, 0.943)	78.0% (69.7%, 84.8%) (99/127)	90.9% (89.4%, 92.2%) (1,609/1,771)	35.1 (22.4, 55.0)
Not Hispanic	43,088	0.902 (0.898, 0.907)	76.2% (74.8%, 77.5%) (3,137/4,118)	87.3% (87.0%, 87.6%) (34,016/38,970)	22.0 (20.3, 23.7)
E. LVEF estimation approach					
3D echocardiography	1,560	0.951 (0.935, 0.967)	82.8% (73.9%, 89.7%) (82/99)	91.6% (90.1%, 93.0%) (1,339/1,461)	52.9 (30.4, 92.1)
Biplane approach using the Simpson meth	20,970	0.904 (0.898, 0.910)	79.0% (77.4%, 80.6%) (2,081/2,633)	85.7% (85.2%, 86.2%) (15,720/18,337)	22.6 (20.4, 25.1)
2–D method	21,126	0.889 (0.880, 0.898)	71.2% (68.9%, 73.5%) (1,048/1,471)	88.3% (87.9%, 88.8%) (17,363/19,655)	18.8 (16.6, 21.2)
M–mode	1,197	0.944 (0.896, 0.991)	70.0% (34.8%, 93.3%) (7/10)	94.4% (93.0%, 95.7%) (1,121/1,187)	39.6 (10.0, 156.8)
Visual estimation	133	0.743 (0.642, 0.844)	56.2% (37.7%, 73.6%) (18/32)	81.2% (72.2%, 88.3%) (82/101)	5.5 (2.4, 13.1)
F. Care setting					
Outpatient	26,763	0.925 (0.919, 0.930)	79.6% (77.7%, 81.4%) (1,524/1,914)	89.2% (88.9%, 89.6%) (22,177/24,849)	32.4 (28.8, 36.5)
Inpatient	5,688	0.866 (0.851, 0.880)	74.7% (71.4%, 77.7%) (566/758)	83.1% (82.0%, 84.2%) (4,098/4,930)	14.5 (12.1, 17.4)
Emergency	2,827	0.897 (0.880, 0.915)	77.9% (73.0%, 82.3%) (254/326)	84.7% (83.2%, 86.1%) (2,118/2,501)	19.5 (14.7, 25.9)
G. Site location					
MN	20,842	0.909 (0.903, 0.915)	78.5% (76.6%, 80.2%) (1,639/2,089)	86.8% (86.3%, 87.3%) (16,280/18,753)	24.0 (21.4, 26.8)
AZ	9,688	0.903 (0.891, 0.914)	75.3% (72.1%, 78.2%) (599/796)	88.6% (87.9%, 89.3%) (7,878/8,892)	23.6 (19.9, 28.1)
FL	9,239	0.906 (0.894, 0.918)	75.3% (71.8%, 78.5%) (497/660)	88.9% (88.2%, 89.6%) (7,628/8,579)	24.5 (20.2, 29.6)
Midwest Health System	5,217	0.869 (0.855, 0.883)	71.6% (68.1%, 74.9%) (501/700)	85.0% (83.9%, 86.0%) (3,839/4,517)	14.3 (11.9, 17.1)
Overall	**44,986**	**0.903 (0.898, 0.908)**	**76.2% (74.9%, 77.5%) (3,236/4,245)**	**87.4% (87.1%, 87.8%) (35,625/40,741)**	**22.3 (20.7, 24.1)**

Figure 2.

Forest plot for artificial intelligence electrocardiogram subgroup performance. Location abbreviations in 2G are as follows: MN-Minnesota, AZ-Arizona, FL-Florida. Odds ratios (ORs) are ‘diagnostic odds ratio’ defined as the ratio between the odds of test positivity in a patient with disease and the odds of test positivity in a patient without disease.



Group	N	AUC	Sensitivity	Specificity	Odds Ratio
A. Normal sinus rhythm					
No	7,874	0.836 (0.825, 0.847)	80.5% (78.4%, 82.4%) (1,266/1,573)	71.1% (70.0%, 72.2%) (4,480/6,301)	10.1 (8.9, 11.6)
Yes	37,047	0.910 (0.904, 0.916)	73.6% (71.8%, 75.2%) (1,950/2,651)	90.5% (90.1%, 90.8%) (31,112/34,396)	26.4 (24.0, 28.9)
B. Paced					
No	41,606	0.902 (0.896, 0.907)	72.7% (71.2%, 74.2%) (2,375/3,266)	89.4% (89.1%, 89.7%) (34,267/38,340)	22.4 (20.6, 24.4)
Yes	3,315	0.817 (0.801, 0.832)	87.8% (85.5%, 89.8%) (841/958)	56.2% (54.2%, 58.2%) (1,325/2,357)	9.2 (7.5, 11.4)
C. Atrial flutter/fibrillation					
No	38,949	0.912 (0.907, 0.917)	76.1% (74.6%, 77.5%) (2,423/3,185)	89.2% (88.9%, 89.6%) (31,916/35,764)	26.4 (24.2, 28.8)
Yes	5,972	0.826 (0.812, 0.840)	76.3% (73.6%, 78.9%) (793/1,039)	74.5% (73.3%, 75.7%) (3,676/4,933)	9.4 (8.1, 11.0)
D. AV Block					
No	38,768	0.905 (0.900, 0.910)	75.6% (74.1%, 77.0%) (2,667/3,528)	88.3% (88.0%, 88.7%) (31,132/35,240)	23.5 (21.6, 25.5)
Yes	6,153	0.886 (0.874, 0.898)	78.9% (75.7%, 81.9%) (549/696)	81.7% (80.7%, 82.7%) (4,460/5,457)	16.7 (13.8, 20.3)
E. Complete heart block					
No	44,844	0.903 (0.898, 0.907)	76.1% (74.8%, 77.4%) (3,208/4,215)	87.5% (87.2%, 87.8%) (35,551/40,629)	22.3 (20.7, 24.1)
Yes	77	0.873 (0.776, 0.969)	88.9% (51.8%, 99.7%) (8/9)	60.3% (47.7%, 72.0%) (41/68)	12.1 (1.4, 102.7)
F. Right bundle branch block					
No	41,275	0.903 (0.899, 0.908)	76.1% (74.7%, 77.5%) (2,933/3,853)	87.6% (87.2%, 87.9%) (32,772/37,422)	22.5 (20.7, 24.3)
Yes	3,646	0.894 (0.877, 0.911)	76.3% (71.6%, 80.5%) (283/371)	86.1% (84.9%, 87.3%) (2,820/3,275)	19.9 (15.4, 25.8)
G. Left bundle branch block					
No	43,158	0.902 (0.897, 0.907)	73.7% (72.2%, 75.1%) (2,710/3,679)	89.0% (88.7%, 89.3%) (35,149/39,479)	22.7 (21.0, 24.6)
Yes	1,763	0.791 (0.769, 0.814)	92.8% (90.3%, 94.9%) (506/545)	36.4% (33.7%, 39.1%) (443/1,218)	7.4 (5.2, 10.5)
H. LV hypertrophy					
No	40,744	0.905 (0.900, 0.910)	75.1% (73.7%, 76.5%) (2,687/3,577)	88.6% (88.3%, 88.9%) (32,937/37,167)	23.5 (21.7, 25.5)
Yes	4,177	0.873 (0.859, 0.887)	81.8% (78.6%, 84.7%) (529/647)	75.2% (73.8%, 76.6%) (2,655/3,530)	13.6 (11.0, 16.8)
I. Infraction present					
No	37,579	0.907 (0.902, 0.913)	76.0% (74.5%, 77.5%) (2,505/3,296)	88.4% (88.0%, 88.7%) (30,293/34,283)	24.0 (22.1, 26.2)
Yes	7,342	0.877 (0.866, 0.888)	76.6% (73.8%, 79.3%) (711/928)	82.6% (81.7%, 83.5%) (5,299/6,414)	15.6 (13.2, 18.4)
J. Low voltage					
No	42,975	0.904 (0.899, 0.909)	76.4% (75.1%, 77.7%) (3,049/3,990)	87.5% (87.2%, 87.8%) (34,114/38,985)	22.7 (21.0, 24.6)
Yes	1,946	0.879 (0.855, 0.902)	71.4% (65.1%, 77.1%) (167/234)	86.3% (84.6%, 87.9%) (1,478/1,712)	15.7 (11.5, 21.6)
K. Prolonged QT					
No	41,330	0.904 (0.898, 0.909)	74.0% (72.5%, 75.5%) (2,485/3,356)	88.8% (88.4%, 89.1%) (33,703/37,974)	22.5 (20.7, 24.5)
Yes	3,591	0.848 (0.833, 0.862)	84.2% (81.6%, 86.6%) (731/868)	69.4% (67.6%, 71.1%) (1,889/2,723)	12.1 (9.9, 14.8)
L. Completely normal ECG					
No	38,359	0.892 (0.887, 0.897)	76.6% (75.2%, 77.8%) (3,210/4,193)	85.2% (84.8%, 85.6%) (29,119/34,166)	18.8 (17.4, 20.4)
Yes	6,562	0.803 (0.733, 0.873)	19.4% (7.5%, 37.5%) (6/31)	99.1% (98.9%, 99.3%) (6,473/6,531)	26.8 (10.6, 67.7)
Overall	**44,986**	**0.903 (0.898, 0.908)**	**76.2% (74.9%, 77.5%) (3,236/4,245)**	**87.4% (87.1%, 87.8%) (35,625/40,741)**	**22.3 (20.7, 24.1)**

1 10 100 1000

Figure 3.

Forest plot for artificial intelligence electrocardiogram performance with respect to electrocardiogram features. Odds ratios (ORs) are ‘diagnostic odds ratio’ defined as the ratio between the odds of test positivity in a patient with disease and the odds of test positivity in a patient without disease.

Table 1

Baseline cohort characteristics with and without left ventricular systolic dysfunction

Age (95% CI)	63.6 (63.4, 63.7)	68.5 (68.0, 69.0)
Female (%)	18270 (44.5)	1082 (27.5)
Race (%)		
Non-Hispanic White	36601 (89.1)	3513 (89.4)
Black	1810 (4.4)	206 (5.2)
Other	2647 (6.4)	209 (5.3)
Hispanic (%)	1776 (4.3)	121 (3.1)
Congestive heart failure (%)	14530 (35.4)	3776 (95.9)
Myocardial infarction (%)	6498 (15.8)	1625 (41.4)
Hypertension (%)	27090 (66.0)	3066 (78.1)
Diabetes mellitus (%)	10283 (25.0)	1398 (35.6)
Renal disease (%)	11997 (29.2)	1806 (46.0)
Cerebrovascular disease (%)	8645 (21.1)	1055 (26.9)
Peripheral vascular disease (%)	16947 (41.3)	2611 (66.5)
COPD (%)	11613 (28.3)	1380 (35.1)
Connective tissue/rheumatologic disease (%)	2963 (7.2)	284 (7.2)
EF below 30% by TTE (%)	0 (0.0)	1848 (47.0%)
EF below 50% by TTE (%)	4262 (10.4)	3928 (100%)

*CI, confidence interval; EF, ejection fraction; TTE, transthoracic echocardiogram