



Artificial Intelligence on the Twelve-Lead Electrocardiogram: What It Detects, What It Cannot, and Who Is Accountable

Wiebke Stollenberg^{1*}, Filipe Rocha-Salgueiro², Nadja Petrovic-Zorko³

¹Centre for Clinical Machine Learning, Ruhr Institute of Cardiovascular Medicine, Essen, Germany

²Department of Cardiology, Vouga Heart Centre, Aveiro, Portugal

³Centre for Medical Law and Ethics, Sava Heart Centre, Ljubljana, Slovenia

*Correspondence to: Wiebke Stollenberg, Centre for Clinical Machine Learning, Ruhr Institute of Cardiovascular Medicine, Essen, Germany

E-mail: w.stollenberg@ricm.de

Received: 04 January 2022; Accepted: 04 June 2022; Published: 27 October 2022

Copyright: © 2022 Stollenberg W, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

A deep-learning model trained on the standard twelve-lead electrocardiogram (ECG) can now detect a left ventricular ejection fraction (EF) of 35% or below directly from the waveform, at an area under the receiver-operating-characteristic curve (AUC) of 0.93 in an independent test cohort of 52,870 patients, with sensitivity 86.3% and specificity 85.7%. In a subsequent bias assessment, that accuracy held up across every racial and ethnic subgroup tested (AUC 0.930-0.961), even though a separate network could distinguish Black from white patients from ECG features alone -- a reassuring finding on a question the field was right to ask. When deployed in a pragmatic, cluster-randomised trial across 45 primary-care clinics (the EAGLE trial), the algorithm increased low-EF diagnosis rates from 1.6% to 2.1% overall, and from 14.5% to 19.5% among patients it flagged as high-risk -- a rare instance of an ECG-based artificial intelligence (AI) tool demonstrating a real-world clinical benefit under randomised conditions rather than only strong retrospective accuracy. Nearly 90 FDA-authorised AI/ML cardiovascular devices now exist, out of 882 across all specialties. This review sets out that evidence and closes with the complication its subtitle names directly: the legal and regulatory frameworks that determine who is accountable when an AI-ECG interpretation is wrong have not kept pace with the technology's clinical deployment, and current liability law concentrates responsibility on the treating physician largely regardless of whether the algorithm's error was reasonably detectable at the bedside.

Keywords: Artificial intelligence; Electrocardiography; Deep learning; Ejection fraction; Medical liability

Introduction

The twelve-lead electrocardiogram (ECG) is the most widely performed cardiac investigation in medicine, and for a century its interpretation has depended on visible, named waveform features -- intervals, axes, ST-segment deviations -- that a trained clinician can point to and explain. Deep-learning models trained directly on the raw ECG waveform have upended that premise: a convolutional neural network can now detect conditions from the ECG that no clinician can read from

it directly, most strikingly a left ventricular ejection fraction (EF) at or below 35%, inferred from waveform patterns the model has learned but that were never part of any clinician's diagnostic vocabulary [1]. This review sets out what these AI-ECG models have been shown to detect, in trial conditions and in a genuine randomised deployment; what a rigorous bias assessment has and has not found; and closes with the complication this review's subtitle asks directly -- who is accountable when the model is wrong, given that the clinician using it typically cannot independently verify its reasoning at

the point of care.

Low EF is not the only condition an AI-ECG model has been shown to detect from features invisible to a human reader: a related deep-learning algorithm can identify patients with paroxysmal atrial fibrillation from a single ECG recorded while they are still in normal sinus rhythm, at an AUC of 0.87 in a retrospective outcome-prediction analysis, again from waveform features that carry no established name in conventional ECG interpretation [4]. Figure 1 sets out the discriminative performance behind the low-EF claim as an illustrative receiver-operating-characteristic curve, anchored to the reported area-under-curve (AUC) statistics from the model's original validation and its strongest-performing demographic subgroup, described further below.

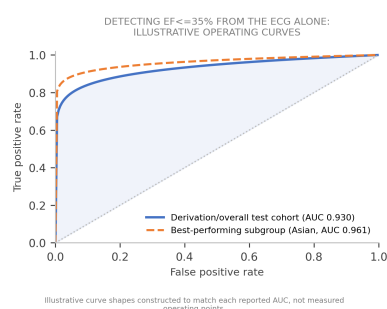


Figure 1: Illustrative receiver-operating-characteristic curves for AI-ECG detection of ejection fraction $\leq 35\%$, anchored to the reported overall and best-subgroup AUC values [1,2].

Material and Methods

This is a narrative review of the primary validation, bias-assessment, and randomised-trial literature on deep-learning interpretation of the standard twelve-lead ECG, together with regulatory documentation and health-law commentary on AI accountability, identified through PubMed, the FDA's public AI/ML device listing, and citation tracking of the sources above, without a pre-registered protocol or systematic search string. Diagnostic-accuracy and trial effect estimates are reported exactly as their source studies present them; no numerical value in this review has been invented, estimated, or extrapolated beyond what the cited source itself reports.

Results

A model trained only on the ECG waveform detects a hidden reduction in cardiac pumping function

Attia and colleagues trained a convolutional neural network on 44,959 patients with a paired ECG and echocardiogram, then tested it on an independent cohort of 52,870 patients, to detect an ejection fraction at or below 35% -- the threshold used clinically to define significant left ventricular systolic dysfunction -- from the ECG waveform alone [1]. In the independent test cohort, the model achieved an AUC of 0.93, with sensitivity 86.3% and specificity 85.7% [1]. A further, clinically important finding concerned patients whose EF was normal at the time of testing: among those without ventricular dysfunction at baseline, an AI-ECG-positive result carried a hazard ratio of 4.1 (95% CI 3.3-5.0) for developing low EF in the future, compared with patients the model classified as negative [1]. Figure 2 shows what happened when this model was subsequently tested across racial and ethnic subgroups it had not been specifically trained to balance for.

A bias assessment found the model's accuracy held up across every racial and ethnic subgroup tested

Because a model trained predominantly on one demographic group can under-perform on others, Attia and colleagues conducted a dedicated fairness assessment of the same low-EF model [2]. AUC was 0.931 in the non-Hispanic white cohort (n=44,524, the group that dominated model training), 0.961 in the Asian subgroup (n=557), 0.937 in the Black/African American subgroup (n=651), 0.937 in the Hispanic/Latino subgroup (n=331), and 0.938 in the American Indian/Native Alaskan subgroup (n=223) [2]. Every subgroup, including the smallest, scored at or above the overall cohort's AUC. This finding is notable precisely because it was not guaranteed: a separately trained network, given the same ECGs, could distinguish Black from white patients with an AUC of 0.84 and 0.76 respectively, confirming that the ECG waveform itself does carry detectable racial signal [2]. That the dysfunction-detection model's accuracy did not degrade despite this underlying signal suggests the features it uses to detect low EF are functionally independent of the features that happen to correlate with race -- a reassuring result on a question this field was right to ask, though one confirmatory study in a single health system does not settle the question for every subsequent AI-ECG model or population. An accompanying editorial commentary on this bias assessment cautioned that a fairness finding specific to one model and one outcome should not be generalised into a blanket assurance that AI-ECG tools are bias-free, and called for bias assessment to become a routine, reported part of validating any clinical AI model rather than an occasional add-on study [6].

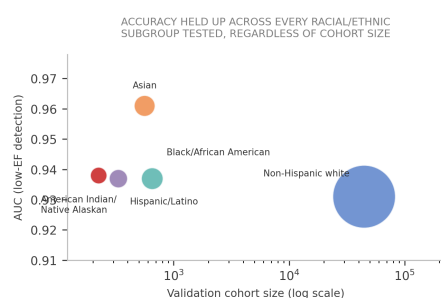


Figure 2: AUC for AI-ECG detection of ejection fraction $\leq 35\%$ by racial/ethnic subgroup, against each subgroup's validation cohort size [2].

The EAGLE trial shows the algorithm changes what happens in routine primary care, not just what a retrospective model can detect

Retrospective accuracy is not the same as real-world clinical benefit, and the EAGLE trial tested the difference directly: a pragmatic, cluster-randomised trial across 45 primary-care clinics (120 care teams, 181 clinicians in the intervention arm and 177 in control), covering 22,641 patients, in which intervention-arm clinicians received the AI-ECG's low-EF risk flag at the point of care and control-arm clinicians did not [3]. Low-EF diagnosis rates rose from 1.6% in the control arm to 2.1% in the intervention arm overall (odds ratio 1.32, $P=.007$), and, among the 6% of the cohort the algorithm flagged as high-risk, from 14.5% to 19.5% (odds ratio 1.43, $P=.01$) [3]. Echocardiogram ordering was similar between arms overall (18.2% versus 19.2%, $P=.17$), but among AI-ECG-positive patients specifically, ordering rose sharply, from 38.1% to 49.6% ($P<.001$) [3] -- clinicians used the flag to direct confirmatory testing where the algorithm suggested it mattered most, exactly the intended use pattern. Figure 3 sets out all four of these comparisons together. Table 1 places the EAGLE trial and the AI-ECG literature it extends alongside the growth of the broader regulatory category this technology now occupies.

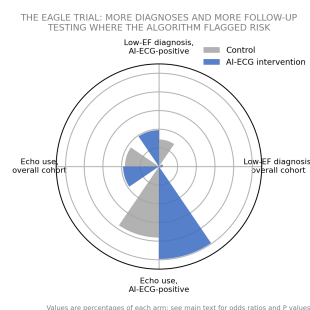


Figure 3: Low ejection fraction diagnosis and follow-up echocardiogram utilisation, intervention versus control, EAGLE pragmatic cluster-randomised trial [3].

A separate, longer-horizon finding from the same model rounds out this evidence: Figure 4 returns to the hazard ratio noted above for patients whose EF was normal at baseline but whose AI-ECG was positive, since it bears directly on how a clinician should act on a positive result in a patient who looks well today.

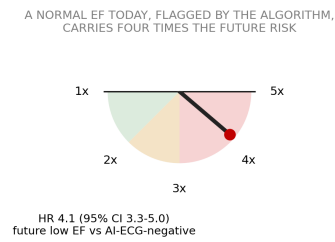


Figure 4: Hazard ratio for future development of low ejection fraction among patients with a normal ejection fraction at baseline but an AI-ECG-positive result [1].

Study	Design	Key result
Attia 2019 (derivation)	Retrospective, N=97,829 (train+test)	AUC 0.93, sens 86.3%, spec 85.7% [1]
Attia 2021 (bias)	Retrospective, 5 subgroups, N~46,286	AUC 0.930-0.961 across subgroups [2]
Yao/EAGLE 2021	Pragmatic cluster-RCT, N=22,641	Diagnosis OR 1.32 overall, 1.43 high-risk [3]

Table 1: AI-ECG low ejection fraction evidence base, from initial validation through a pragmatic randomised deployment trial [1,2,3].

The technology has scaled faster than the legal framework for what happens when it errs

The clinical evidence above sits inside a rapidly growing regulatory category: as of March 2024, the FDA had authorised 882 AI/ML-enabled medical devices across all specialties, of which 90 were cardiovascular applications, a subset that has grown at roughly 75% per year since the first such device was cleared in 2015 [5]. The FDA itself has acknowledged that its traditional device-regulation paradigm, built around a fixed product cleared once, was not designed for adaptive AI/ML software that can continue to learn after deployment, and its 2021 AI/ML Software as a Medical Device Action Plan and subsequent guidance on good machine-learning practice and predetermined change-control plans represent an explicit, still-evolving attempt to build a lifecycle-based regulatory framework for exactly this category of device [8]. Figure 5 sets that cardiovascular share against the wider device category. The complication this review closes on is not a question of whether these algorithms work in the specific, narrow senses the evidence above documents -- they plainly do -- but of what happens, legally and practically, when one of them is wrong in a way that harms a patient, and who bears that consequence. Current US liability analysis finds that physicians retain primary legal responsibility for a diagnostic decision regardless of an AI/ML algorithm's output, and courts have historically rejected physician attempts to shift blame to external sources such as device manufacturers' documentation [7]. Manufacturer and developer liability, by contrast, remains legally unsettled: a products-liability claim against an algorithm developer requires the plaintiff to show a viable, cost-effective alternative design existed, a standard that is difficult to apply to a trained neural network, and courts have ruled inconsistently on comparable device- and software-liability cases [7]. Health systems sit in between, potentially liable through negligent credentialing of an AI tool before deployment, vicarious liability for a clinician's error while using it, or direct negligence in training or maintenance [7]. Figure 6 sets this current distribution out as a chain running from the patient to the developer.

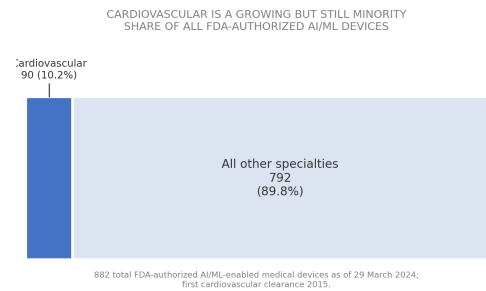


Figure 5: FDA-authorised AI/ML-enabled medical devices by specialty, cardiovascular versus all others, as of March 2024 [5].

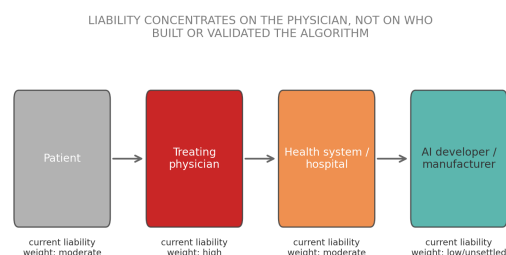


Figure 6: Current distribution of legal liability across the stakeholders in an AI-ECG diagnostic error, from patient to algorithm developer [7].

Discussion

The evidence reviewed here supports a specific, qualified form of enthusiasm: AI-ECG interpretation for low ejection fraction is not merely an impressive retrospective statistic -- it is one of a small number of AI-based cardiovascular tools with a positive pragmatic randomised trial behind it, and its accuracy does not show the demographic degradation that critics of medical AI have reasonably worried about in other applications [2,3]. That combination -- prospective clinical benefit plus a documented absence of the specific bias tested for -- is a higher evidentiary bar than most of the 882 currently authorised AI/ML medical devices have cleared, since the great majority reach market on retrospective validation alone. The accountability gap this review describes is therefore not a reason to distrust this specific algorithm's outputs; it is a reason to recognise that the legal infrastructure surrounding its use has not caught up with either its accuracy or its demonstrated clinical benefit. A physician using AI-ECG screening in the way EAGLE validated it -- as a risk-stratification flag prompting confirmatory echocardiography, not as a stand-alone diagnosis -- is following the trial's actual protocol and remains professionally responsible for that clinical judgment, exactly as current liability law expects. The harder case, not yet tested at scale in courts, is the AI-ECG-negative patient who is not flagged, does not receive confirmatory testing, and is later found to have had low EF at the time of the ECG that the algorithm read as normal: the physician had no waveform feature to independently notice, no comparably accurate alternative reading method, and yet, under the framework described above, likely remains the principal locus of liability.

The Milbank Quarterly analysis reviewed here proposes that liability approaches should differentiate between transparent, static algorithms and opaque, self-updating ones, treating the latter as candidates for specialised no-fault compensation schemes rather than conventional negligence litigation [7]. The AI-ECG low-EF model, as validated and deployed in EAGLE, sits closer to the transparent, static end of that spectrum -- it is a fixed, published, externally validated classifier, not a continuously self-updating system -- which arguably makes conventional liability frameworks more

workable for it than for more adaptive AI/ML devices entering cardiovascular practice behind it. But that argument has not yet been tested in a reported case, and the growth rate of AI/ML cardiovascular device authorisations documented here means the number of devices for which this question will eventually need answering is expanding well ahead of any legal or regulatory resolution.

Study limitations

This is a narrative review, not a systematic review with a pre-registered protocol, and its clinical evidence base concentrates on a single research group's AI-ECG low-EF model and a single pragmatic trial of it; other AI-ECG applications (atrial fibrillation, hypertrophic cardiomyopathy, and hyperkalaemia detection among them) have their own, separately reported evidence bases not exhaustively surveyed here. The bias assessment discussed reflects one model, one health system's patient population, and one set of demographic categories; it should not be read as establishing that AI-ECG models are free of demographic bias in general. The liability analysis in this review reflects US legal commentary and case law as of the cited source's publication; liability frameworks differ materially by jurisdiction and continue to evolve as courts encounter AI-specific cases.

Conclusion

Artificial intelligence applied to the twelve-lead ECG can detect a clinically important, otherwise hidden reduction in cardiac pumping function with an accuracy that held up in an independent test cohort of over 50,000 patients, held up again across every racial and ethnic subgroup a dedicated bias assessment tested, and translated into a measurable increase in real-world diagnosis rates when deployed in a genuine pragmatic randomised trial across dozens of primary care clinics. That is a substantially stronger evidentiary record than most of the hundreds of AI/ML devices now authorised across medicine can claim, and it justifies describing this specific application as a genuine advance rather than a promising but unproven one. What the evidence does not yet resolve is the question this review's title asks directly: when an AI-ECG interpretation is wrong, current liability law places the practical burden of that error almost entirely on the treating physician, largely independent of whether the error was one a clinician using the tool as intended could reasonably have caught. Closing that gap does not require slowing the technology's clinical deployment, which the evidence here supports; it requires the legal and regulatory frameworks governing it to be built with the same rigour that produced the clinical evidence itself.

Until that legal work is done, the most defensible position for a clinician using AI-ECG tools is the one EAGLE's own protocol modelled: treat the algorithm's output as a

risk-stratification input that changes the probability of ordering a confirmatory test, not as a diagnosis in itself, and document that reasoning at the point of care. That practice does not resolve the underlying accountability question this review has described, but it is the closest available match, today, between how these tools have actually been shown to work and how liability law currently expects clinicians to use them.

1. Attia ZI, Kapa S, Lopez-Jimenez F, et al. "Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram". *Nat Med* 25(2019):70-74.
2. Attia ZI, Harmon DM, Behnken EM, 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* 14(2021):e007988.
3. Yao X, Rushlow DR, Inselman JW, et al. "Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial". *Nat Med* 27(2021):815-819.
4. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction". *Lancet* 394(2019):861-867.
5. Agarwal N. "AI in MedTech #2: Insights from 882 device approvals through March 2024" [newsletter]. Accessed 2026.
6. Ho SS, Bierer BE, et al. "Assessing and Mitigating Bias in Medical Artificial Intelligence" [editorial commentary]. *Circ Arrhythm Electrophysiol* 14(2021):e010154.
7. Milbank Memorial Fund. "Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation". *Milbank Q* (2024).
8. U.S. Food and Drug Administration. "Artificial Intelligence in Software as a Medical Device". FDA AI/ML SaMD Action Plan, January 2021, and subsequent guidance.

Cite this article: Stollenberg W, Rocha-Salgueiro F, Petrovic-Zorko N. Artificial Intelligence on the Twelve-Lead Electrocardiogram: What It Detects, What It Cannot, and Who Is Accountable. *J Heart Cardiovasc Sci.* 2022; 3: 007-012.