WISELEAF SCIENTIFIC VENTURES

SMART HEALTHCARE REVOLUTION: AI STRATEGIES FOR SUPERIOR GLOBAL HEALTH

CHAPTER 2- AI-POWERED CARDIOVASCULAR RISK PREDICTION

Mr. Praful Sunil Urade

Assistant Professor, Department of Pharmaceutical Science, Manwatkar College of Pharmacy, Ghodpeth, Chandrapur, Maharashtra, India

Email: Prafulurade.mcpc@gmail.com

ABSTRACT

Cardiovascular disease remains the leading cause of death worldwide, and most of those deaths now occur in low- and middle-income countries, India prominent among them. Prevention depends on identifying people at high risk before the first event, a task long performed by regression-based scores such as Framingham, SCORE2, QRISK3, the Pooled Cohort Equations and the WHO risk charts. These instruments are transparent and inexpensive, but they were derived mostly in European-ancestry populations and tend to misjudge risk in South Asians, who develop coronary disease earlier. This chapter examines what artificial intelligence (AI) has added to cardiovascular risk prediction and where its claims outrun the evidence. Machine learning applied to electronic health records and large cohorts has produced modest but consistent gains in discrimination, with C-statistics rising from about 0.72 to 0.76-0.77 in British primary care and UK Biobank data. The larger change has come from new inputs. Deep learning applied to the standard 12-lead electrocardiogram can detect left ventricular systolic dysfunction and signal hidden atrial fibrillation, and it is among the few cardiovascular AI tools tested in pragmatic randomised trials, where it raised the yield of new diagnoses in primary care and hospital settings. Smartwatch rhythm notification, retinal photographs and coronary imaging add further channels, with more uneven evidence. The chapter then turns to explainability, calibration, ethnic bias and the frequent loss of performance at external validation, and summarises recent regulatory authorisations. A final section considers implementation within India's National Programme for Prevention and Control of Non-Communicable Diseases (NP-NCD) and sets out priorities for locally derived, prospectively evaluated tools.

Keywords: cardiovascular risk prediction; machine learning; AI-enabled electrocardiography; calibration; South Asians; external validation

1. Introduction

Cardiovascular disease (CVD) caused an estimated 19.2 million deaths worldwide in 2023, about one in three of all deaths, up from 13.1 million in 1990, according to the latest Global

Burden of Disease (GBD) analysis [1]. An earlier WHO analysis estimated that roughly three-quarters of CVD deaths occur in low- and middle-income countries [2]. India illustrates the shift sharply. In 2016 ischaemic heart disease was the leading single cause of disability-adjusted life years (DALYs) in the country, and cardiovascular diseases accounted for 14.1% of all DALYs and about 28% of all deaths [3]. Cardiovascular conditions make up about two-thirds of India's non-communicable disease burden [4,5]. The international INTERHEART comparison found that the mean age at first myocardial infarction was 53.0 years in South Asian countries against 58.8 years elsewhere [6], and the CREATE registry of 20,468 patients with acute coronary syndromes across 89 Indian centres showed that poorer patients were less likely to receive evidence-based treatments such as lipid-lowering drugs and percutaneous intervention, and had higher 30-day mortality (8.2% vs 5.5%) [7].

Most of this burden is preventable in principle. Blood pressure lowering, statins, smoking cessation and glycaemic control all reduce events, but they must be directed at the right people, and the clinical tool for that direction is the risk score. A risk score converts a handful of measurements into an absolute probability of an event over a fixed horizon, usually ten years, and the probability is compared with a treatment threshold. The logic is simple. The difficulty lies in whether the probability is correct for the person in front of the clinician.

Artificial intelligence (AI), in the narrow sense of statistical learning from large datasets, approaches this problem in two different ways. The first treats risk prediction as an exercise in better curve fitting: the same variables, or many more of them, are fed into flexible algorithms such as gradient-boosted trees or neural networks in the hope of capturing interactions and non-linearities that a Cox model misses. The second uses AI to read signals that clinicians cannot easily quantify, such as the raw voltage of the electrocardiogram (ECG), the pattern of retinal vessels, or the texture of coronary plaque on computed tomography (CT). The first approach has yielded incremental gains. The second has produced some of the most clinically interesting results in the field, and it is also where the risk of overclaiming is greatest [8].

This chapter follows that division. Section 2 reviews the conventional scores and their weaknesses in South Asian populations. Section 3 considers machine learning on routinely collected and cohort data. Section 4 covers AI-enabled electrocardiography, including the randomised evidence. Section 5 turns to wearables, Section 6 to imaging and genomic inputs, and Section 7 to explainability, calibration and bias. Section 8 summarises regulatory status, Section 9 discusses implementation in India, and Section 10 sets out future directions. Figure 2.1 gives a schematic of the pipeline that links these parts.

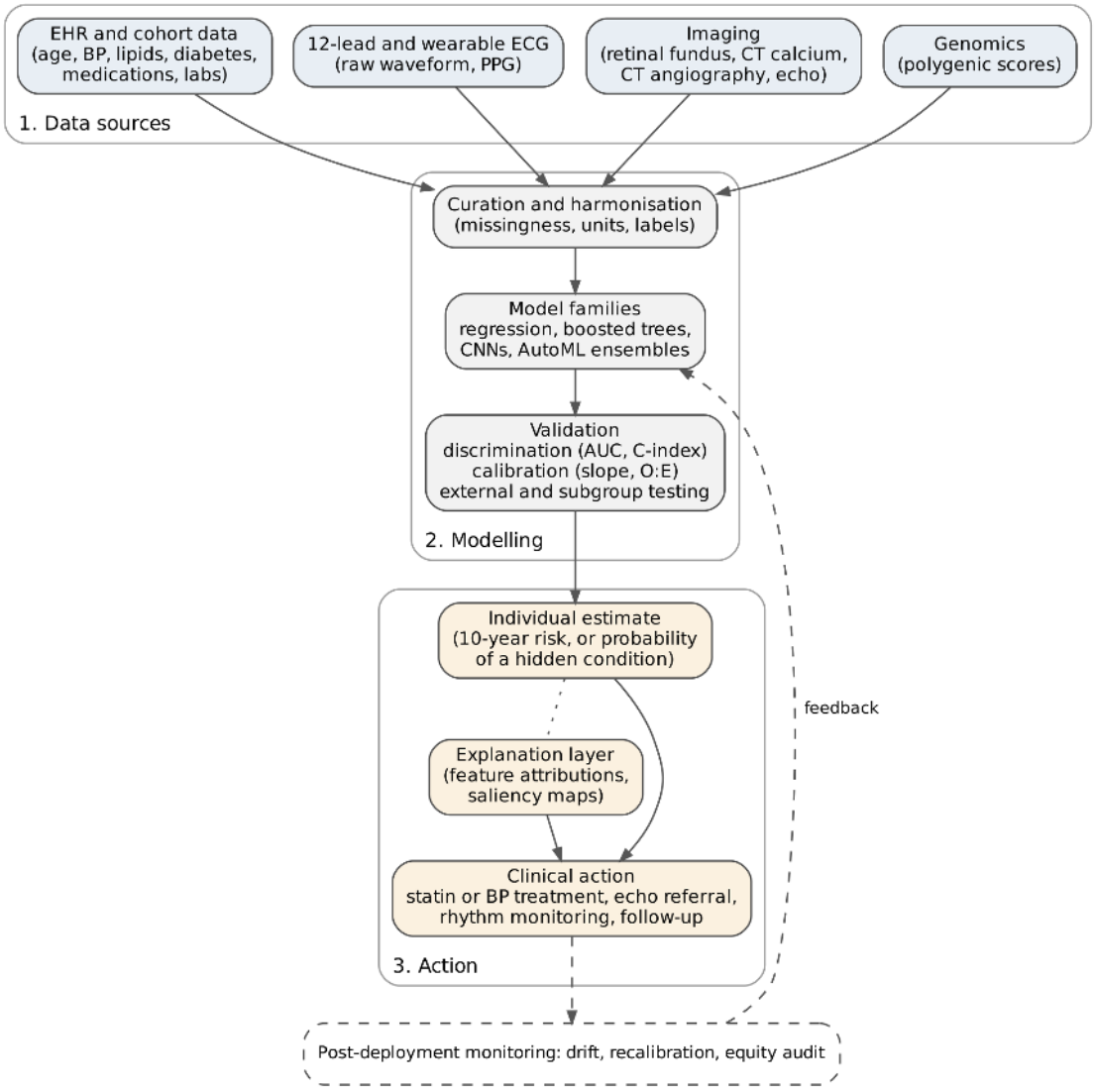

Figure 2.1: Generic pipeline for AI-based cardiovascular risk prediction, from multimodal data through model development and validation to an individual estimate and clinical action, with post-deployment monitoring feeding back into the model. Original schematic prepared for this chapter.

2. Conventional Risk Scores and Their Limits

2.1 From Framingham to the pooled cohort equations

The Framingham Heart Study gave cardiology its first multivariable risk functions. A widely used 1998 version estimated coronary heart disease risk from categories of age, blood pressure, cholesterol, diabetes and smoking [9]. A decade later D'Agostino and colleagues published a general CVD risk profile for primary care, derived in 8,491 Framingham participants aged 30

to 74 years who experienced 1,174 first cardiovascular events over 12 years; its C-statistic was 0.763 in men and 0.793 in women [10]. These functions have been copied into countless calculators and guidelines, and they are the reference point against which almost every later model, including the AI models discussed below, has been judged.

In the United States, the 2013 American College of Cardiology and American Heart Association (ACC/AHA) guideline introduced the Pooled Cohort Equations (PCE), which pooled several community cohorts and produced separate equations for white and African American men and women [11]. Concerns about overestimation in contemporary populations, and about the use of race as an input, informed the AHA's development of the PREVENT equations. PREVENT was derived in 3,281,919 participants from 25 datasets and externally validated in 3,330,085 participants from 21 further datasets; it omits race, incorporates kidney function, and reported median external C-statistics for total CVD of 0.794 in women and 0.757 in men, with calibration slopes close to 1 [12].

2.2 European and British scores

Europe followed a different route. The original SCORE project estimated ten-year risk of fatal CVD from pooled European cohorts [13]. Its successor, SCORE2, was derived in 677,684 individuals from 45 cohorts in 13 countries (30,121 events) and validated in 1,133,181 individuals from 25 cohorts in 15 countries; it predicts fatal and non-fatal events and is recalibrated to four risk regions defined by national CVD mortality, with external C-indices ranging from 0.67 to 0.81 [14]. In the United Kingdom, QRISK3 was developed in 7.89 million patients from 981 general practices and validated in 2.67 million patients from another 328; it added chronic kidney disease, blood pressure variability, migraine, corticosteroid use, systemic lupus erythematosus, atypical antipsychotics and severe mental illness, and it reported Harrell's C of 0.88 in women and 0.86 in men [15]. QRISK3 is unusual in modelling ethnicity directly, with separate categories for Indian, Pakistani, Bangladeshi and other Asian groups [15].

The high C-statistics for QRISK3 deserve comment. They partly reflect the very wide age range of a whole primary-care population, since age alone separates people well over ten years. A C-statistic is not portable: the same model will look weaker in a narrower or older population, which is one reason comparisons across papers must be read with care.

2.3 The WHO risk charts

For settings without cohort data, the World Health Organization (WHO) issued risk charts in 2007 and revised them in 2019. The revised charts were derived in 376,177 individuals from 85 cohorts and recalibrated to 21 Global Burden of Disease regions, including South Asia, using regional incidence of myocardial infarction and stroke and national risk-factor data; external C-indices across 19 validation cohorts ranged from 0.685 to 0.833 [2]. Two versions exist: a laboratory-based chart using cholesterol and diabetes status, and a non-laboratory chart

that substitutes body mass index, which matters in primary health centres where a lipid profile may not be available [2]. The companion WHO HEARTS technical package supplies simplified treatment protocols for CVD management in primary care [16].

2.4 Why South Asians are poorly served

Every score in Table 2.1 was built mainly from European-ancestry or North American data, and South Asians are consistently the group in which they perform worst. In the UK Biobank, 8,124 participants of South Asian ancestry had a twofold higher adjusted hazard of atherosclerotic CVD than 449,349 participants of European ancestry (hazard ratio 2.03), yet the PCE predicted only a 1.1-fold and QRISK3 a 1.4-fold difference in risk [17]. Older work told the same story. In the Newcastle Heart Project, SCORE predicted comparatively low risk in all South Asian groups, and the agreement between predicted and observed mortality varied by subgroup, failing for Indian men [18]. In the tri-ethnic SABRE cohort, both QRISK2 and Framingham underestimated risk in South Asian women, with predicted-to-observed ratios of 0.52 and 0.43 respectively [19].

Within India, the problem is compounded by the near absence of large prospective cohorts with adjudicated outcomes. Recalibration has therefore relied on risk-factor means and national mortality rates rather than locally observed events. When Gupta and colleagues recalibrated Framingham with data from the CARRS and ICMR urban surveys, the share of women classified at above 30% ten-year risk rose from 3.5% with the original model to 16.4% with the CARRS-recalibrated version, and the two local recalibrations disagreed with each other [5]. An earlier recalibration for a rural Indian population showed the same sensitivity to inputs: the mean ten-year coronary risk for men was 10.4% with the original Framingham equation and 10.7% after recalibration to local data, but 18.9% after recalibration to national mortality data [20]. Clinical series point the same way. In a retrospective study of patients presenting with a first acute coronary syndrome at an Indian tertiary hospital, 63.8% had been in the low-risk Framingham category and only 4.8% in the high-risk category before their event [21]. A score that labels most future patients as low risk cannot guide prevention, however well it discriminates on paper.

A further difficulty is the ten-year horizon itself. Because age dominates every score, a 35-year-old with diabetes, central obesity and untreated hypertension will usually fall below a ten-year treatment threshold, even though their cumulative risk over the following three decades is high. This matters more in South Asia than in Europe, because the disease appears earlier [6]. Lifetime or 30-year estimates are one partial answer. Another is to treat South Asian ancestry itself as a reason to lower the threshold for treatment. Neither solves the underlying problem that the equations were not fitted to South Asian outcomes.

Table 2.1: Conventional cardiovascular risk scores compared

Score (year)	Derivation population	Outcome and horizon	Inputs beyond age, sex, BP, smoking	Reported performance	Known issues in South Asians
Framingham general CVD (2008) [10]	8,491 US adults, Framingham	10-year total CVD	Total and HDL cholesterol, diabetes, BP treatment	C 0.763 (men), 0.793 (women)	Misestimation in UK South Asians [18,19]; most Indian ACS patients classed low risk [21]
ACC/AHA PCE (2013) [11]	Pooled US cohorts	10-year hard ASCVD	Cholesterol, diabetes, race-specific equations	Varies by cohort	Predicted 1.1-fold vs observed ~2-fold excess in UK Biobank [17]
AHA PREVENT (2024) [12]	3.28 million US adults, 25 datasets	Total CVD	Kidney function, optional albuminuria; no race term	External C 0.794 (women), 0.757 (men)	Not yet evaluated in Indian cohorts in the sources reviewed
SCORE2 (2021) [14]	677,684 Europeans, 45 cohorts	10-year fatal and non-fatal CVD	Non-HDL cholesterol; four regional recalibrations	External C 0.67-0.81	No South Asian region
QRISK3 (2017) [15]	7.89 million UK primary-care patients	10-year CVD	Ethnicity, deprivation, CKD, SBP variability, SLE, migraine, mental illness, drugs	C 0.88 (women), 0.86 (men)	Models Indian, Pakistani, Bangladeshi groups, yet predicted only 1.4-fold excess [17]
WHO charts (2019) [2]	376,177 adults, 85 cohorts	10-year fatal and non-fatal CVD	Lab version: cholesterol, diabetes; non-lab: BMI	External C 0.685-0.833	South Asia recalibrated from GBD incidence, not local outcomes

3. Machine Learning on Routine Clinical and Cohort Data

3.1 Primary-care records

The first widely cited test of machine learning against a guideline score used the UK Clinical Practice Research Datalink. Weng and colleagues followed 378,256 adults free of CVD at baseline, of whom 24,970 (6.6%) developed a first cardiovascular event over ten years [22]. The ACC/AHA-style reference algorithm achieved an AUC of 0.728. Random forests reached 0.745, logistic regression 0.760, gradient boosting 0.761 and neural networks 0.764; the best model correctly identified 355 more future cases than the established algorithm [22].

Two features of this result are often lost in retellings. The gains are real but small, around 0.03 to 0.04 in AUC. And logistic regression, given the same expanded set of variables, did nearly as well as the neural network. Much of the improvement came from more data per patient rather than from the algorithm itself. This pattern recurs across the literature and should temper any expectation that a change of algorithm alone will transform prediction.

3.2 Automated machine learning in UK Biobank

Alaa and colleagues pushed the idea further with AutoPrognosis, an automated pipeline that searches across imputation methods, feature processors, classifiers and calibrators and combines them in an ensemble [23]. Using 473 variables in 423,604 UK Biobank participants without baseline CVD, of whom 4,801 had an event within five years, AutoPrognosis reached an AUC of 0.774, against 0.724 for the Framingham score and 0.734 for a Cox model with conventional risk factors [23]. A Cox model with all 473 variables achieved 0.758, so again roughly half of the gain came from additional inputs. The model also identified 368 more cases than Framingham, and among its most informative variables were usual walking pace and self-rated overall health, neither of which appears in any guideline score [23].

UK Biobank is a volunteer cohort of middle-aged British adults, and participants of South Asian ancestry make up less than 2% of it [17]. Strong internal performance there says little about how such a model would behave in an Indian district hospital, where many of the 473 variables would not exist.

3.3 What the comparisons show

Figure 2.2 places the published comparisons side by side. In two large record-based studies, machine learning modestly outperformed the conventional comparator. In two studies that introduced new data types, the deep-learning model merely matched the conventional score on hard outcomes, a point developed in Sections 4 and 6. In 2,994,837 English primary-care patients, a time-varying neural network predicted incident atrial fibrillation with an AUC of 0.827, against 0.725 for CHARGE-AF, and reduced the number needed to screen from 13 to 9 at 75% sensitivity [24]; whether such gains translate into better screening programmes still awaits prospective testing.

The gains in Figure 2.2 should be weighed against what they cost. A Framingham estimate needs seven numbers and a sheet of paper. AutoPrognosis needed hundreds of variables, a tuned software stack and a population that resembles UK Biobank. For most clinics in low-resource settings, the conventional score recalibrated to local data is likely to bring larger benefits than a complex model trained elsewhere.

3.4 What makes a comparison fair

Many published claims that machine learning “outperforms” a risk score do not survive close reading, and a few recurring design choices explain why. The comparator is often an off-the-shelf score applied without recalibration, while the new model is fitted to the very population in which it is tested; the comparison then measures local fitting as much as algorithmic advantage. A fairer design refits a conventional regression model on the same training data and with the same candidate variables, as Weng and colleagues did when their logistic regression nearly matched the neural network [22]. A second issue is the split between development and test data. Random splits of a single database allow the same hospitals, the same coding habits

and often the same calendar period to appear on both sides, so the reported AUC is closer to internal than external validity. Temporal splits, in which the model is trained on earlier years and tested on later ones, and geographic splits across regions or health systems give a more honest estimate [27].

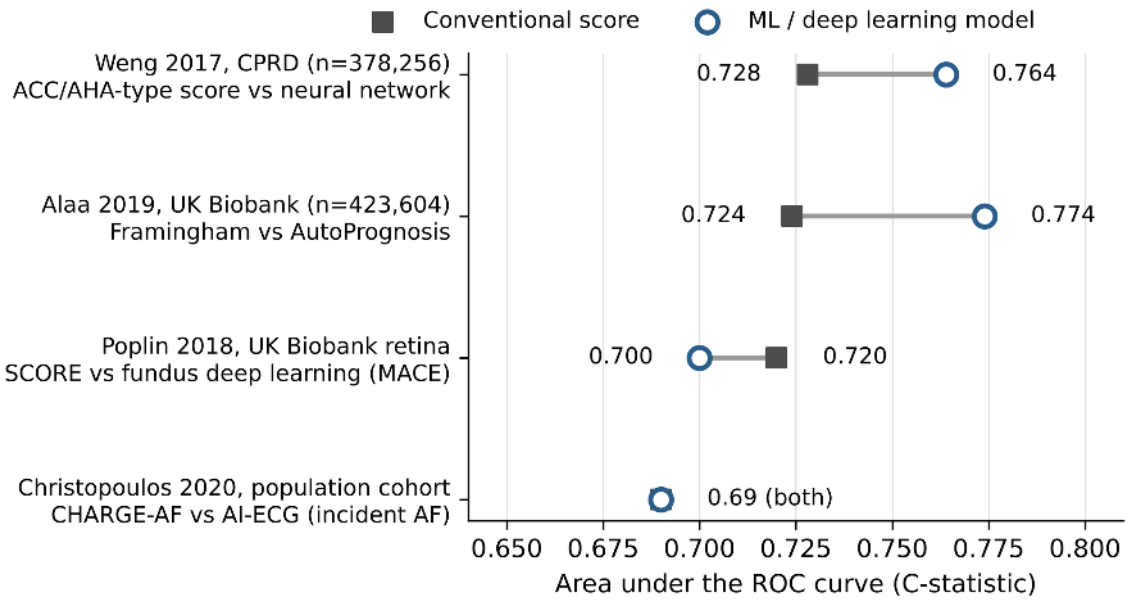

Figure 2.2: Published discrimination (AUC or C-statistic) of conventional scores and machine-learning or deep-learning models evaluated in the same cohorts. Data from Weng et al. [22], Alaa et al. [23], Poplin et al. [25] and Christopoulos et al. [26]. Original chart prepared for this chapter.

Electronic health records add their own hazards. Variables are recorded because a clinician was worried, so the mere presence of a test can carry predictive information that will not transfer to a setting with different ordering habits. Outcomes are captured only when patients return to the same system, and deaths outside hospital may be missed. Missing values are rarely random. Flexible algorithms are good at exploiting all of these patterns, which is precisely why their apparent advantage can vanish in a new setting. None of this argues against machine learning. It argues for judging it by the same standards of external validation and calibration that were eventually applied to QRISK [28].

4. AI-Enabled Electrocardiography

4.1 Reading what cardiologists cannot see

The ECG is cheap, standardised, increasingly digital and performed millions of times a day, which makes it an ideal substrate for deep learning. Convolutional neural networks trained on raw waveforms can recognise rhythm disorders: Hannun and colleagues trained a network on

91,232 single-lead ambulatory recordings from 53,549 patients and reported an average AUC of 0.97 across 12 rhythm classes, with an F1 score of 0.837 against 0.780 for practising cardiologists [29]. That work showed that deep learning could match experts at a task that experts already perform.

The more striking development has been the use of the ECG to infer conditions that are not visible to the human reader. The Mayo Clinic group trained a network on 44,959 paired ECGs and echocardiograms to detect an ejection fraction of 35% or less. In an independent set of 52,870 patients the model achieved an AUC of 0.93, sensitivity 86.3% and specificity 85.7% [30]. A finding with practical weight came from patients whose echocardiogram was normal at the time: those with a positive AI screen had a fourfold higher risk of developing ventricular dysfunction later (hazard ratio 4.1) [30]. A “false positive” may therefore be an early warning rather than an error.

The same group then trained a network on 649,931 normal sinus rhythm ECGs from 180,922 patients to identify people who had, or would soon have, documented atrial fibrillation [31]. From a single sinus rhythm ECG, the AUC was 0.87 and accuracy 79.4%; when all ECGs from the month before a patient’s first recorded episode were used and the highest-risk tracing was taken, the AUC rose to 0.90 and accuracy to 83.3% [31]. The clinical problem behind this work is that paroxysmal atrial fibrillation is intermittent and easily missed, yet it raises stroke risk. A signal that flags who should receive long-term monitoring would make rhythm surveillance far more efficient.

Other targets followed quickly: estimation of age and sex [32], aortic stenosis [33] and heart failure [34]. A model using only two limb leads screened for hyperkalaemia with AUCs of 0.853 to 0.883 across three Mayo Clinic sites [35].

What these networks detect is only partly understood. Some targets have a plausible electrophysiological basis: potassium concentration alters the T wave and QRS morphology, which is why a network can estimate hyperkalaemia from the tracing [35]. Others are less intuitive. The ability to estimate age from the ECG implies that the waveform carries information about cardiac ageing, and a gap between estimated and chronological age has been proposed as a marker of biological ageing [32]. For low ejection fraction, the network appears to integrate many small changes in amplitude, duration and repolarisation that no human reader would weigh together. Because the features are learned rather than specified, it is difficult to say in advance which populations or recording conditions will degrade them, a theme taken up in Section 4.4.

4.2 Predicting future atrial fibrillation

Detecting a hidden condition is not the same as predicting risk, and the distinction matters for prevention. In 1,936 participants of the population-based Mayo Clinic Study of Aging, an AI-

ECG model and the clinical CHARGE-AF score each predicted incident atrial fibrillation with a C-statistic of 0.69, and combining them gave 0.72 [26]. Participants with an AI-ECG output above 0.5 had a cumulative incidence of 21.5% at two years and 52.2% at ten years [26]. In Geisinger data comprising 1.6 million 12-lead ECGs, a deep network predicted new-onset atrial fibrillation within one year with an AUC of 0.85, and it classified as high risk 62% of the patients who went on to have an atrial fibrillation-related stroke within three years [36]. Adding a polygenic score and CHARGE-AF to an ECG model in Montreal Heart Institute data improved model fit but barely changed discrimination, a result discussed in Section 6.3 [37]. These results suggest that the ECG carries prognostic information beyond conventional risk factors. They also show that the advantage over a well-built clinical score can be small when the outcome lies years ahead.

4.3 From accuracy to outcomes: the randomised trials

Most AI studies in cardiology stop at retrospective accuracy. The AI-ECG for low ejection fraction is a rare exception, because it was tested in a pragmatic trial embedded in routine care. The EAGLE trial cluster-randomised 120 primary-care teams in Minnesota and Wisconsin, comprising 358 clinicians and 22,641 adults without known heart failure [38]. In the intervention arm, clinicians received an alert when the AI-ECG was positive, with a recommendation to consider echocardiography. The primary outcome, a new diagnosis of low ejection fraction, occurred in 2.1% of intervention patients and 1.6% of controls (odds ratio 1.32) [38]. Among patients with a positive screen, echocardiography use rose from 38.1% to 49.6%, while overall echocardiography use barely changed [38].

The effect looks modest, and in absolute terms it is: roughly five additional diagnoses per thousand patients screened. Yet the trial showed that an ECG-based alert could change clinician behaviour without flooding echocardiography services, and it was reported in a way that allowed readers to judge exactly what was tested, in line with the CONSORT-AI reporting extension [39]. Even so, about half of the patients with a positive alert did not receive an echocardiogram. That gap between a correct prediction and a completed action is where most of the potential benefit is lost.

The arithmetic of screening explains why. About 6% of patients in each arm of EAGLE had a positive AI-ECG [38]. Even with an AUC above 0.9, the proportion of positive screens that turn out to be true low ejection fraction depends on how common the condition is in the population screened, and in primary care it is uncommon. Among positive-screen patients in EAGLE, a new diagnosis was made in 19.5% of the intervention arm and 14.5% of controls [38]. Put differently, roughly four in five positive alerts did not lead to a diagnosis of low ejection fraction within the trial period, either because the result was a false positive, because echocardiography was not done, or because dysfunction had not yet developed. The Mayo data

suggest that some of these "false positives" will become true positives over the following years [30], but a busy clinician cannot act on that possibility for every alert. Threshold choice, follow-up rules for positive screens with normal echocardiograms, and the capacity of local echocardiography services all need to be designed alongside the algorithm.

A second pragmatic trial, conducted in inpatients at a Taiwanese academic hospital, reached a similar conclusion. Among 13,631 patients, new diagnoses of low ejection fraction within 30 days were more frequent in the AI-ECG arm (1.5% vs 1.1%; hazard ratio 1.50), and the positive predictive value of echocardiograms for low ejection fraction rose from 20.2% to 34.2%, with no change in overall echocardiogram use and no difference in 30-day mortality [40]. Both trials measured diagnostic yield, not hard clinical outcomes. Whether earlier detection of asymptomatic dysfunction translates into fewer hospitalisations or deaths has not yet been shown in a trial. Figure 2.3 summarises the process and yield outcomes of the two trials.

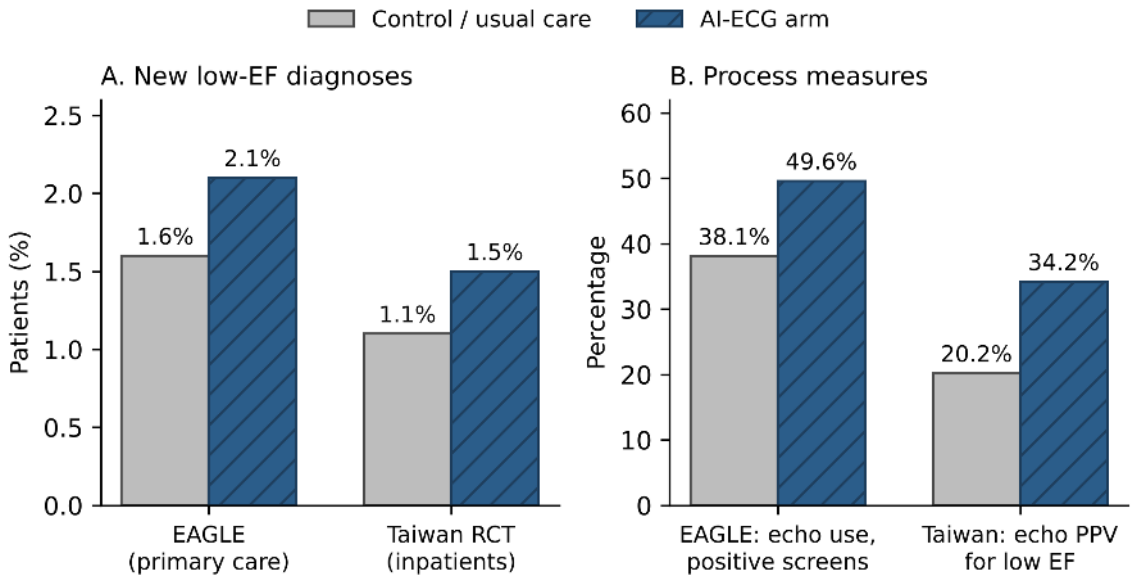

Figure 2.3: Outcomes of the two pragmatic randomised trials of AI-ECG screening for low ejection fraction. Panel A: new diagnoses of low ejection fraction in the control and AI-ECG arms. Panel B: echocardiography use among positive-screen patients (EAGLE) and positive predictive value of echocardiograms for low ejection fraction (Taiwan). Data from Yao et al. [38] and Tsai et al. [40]. Original chart prepared for this chapter.

4.4 Does AI-ECG travel?

External validation of AI-ECG has produced mixed results. When the Mayo low ejection fraction algorithm was applied to 4,277 participants in a Russian population-based study, the AUC fell to 0.82, and at the probability cut-off from the original study sensitivity was only 26.9%, although specificity remained 97.4% [41]. The prevalence of left ventricular systolic

dysfunction was 0.6% in the Russian sample against 7.8% in the clinical derivation cohort [41], so a threshold tuned in one setting was poorly suited to the other. Other external validations have been more reassuring. A Korean aortic stenosis model lost little discrimination, from an internal AUC of 0.884 to 0.861 externally [33], and a previously published model for left ventricular systolic dysfunction, developed by another group, reached an AUC of 0.88 in 42,291 ECG-echocardiogram pairs at a German heart centre, comparable to its published performance [42].

The lesson is not that external validation always degrades performance, but that its direction cannot be predicted in advance. An external cohort drawn from tertiary hospitals with sicker patients and clearer disease can flatter a model; a community population with low prevalence can expose its weaknesses. Only testing in the intended population, with the intended threshold, settles the matter.

Table 2.2: Key studies of AI for cardiovascular risk and hidden disease

Study	Data source and setting	n	Model	Principal metric
Weng 2017 [22]	UK primary-care records (CPRD)	378,256	RF, LR, GBM, neural network	AUC 0.764 (NN) vs 0.728 (ACC/AHA)
Alaa 2019 [23]	UK Biobank, 473 variables	423,604	AutoPrognosis ensemble	AUC 0.774 vs 0.724 (Framingham)
Hannun 2019 [29]	Single-lead ambulatory ECG	53,549 patients	Deep CNN	Mean AUC 0.97; F1 0.837 vs 0.780 (cardiologists)
Attia 2019a [30]	Mayo Clinic ECG-echo pairs	52,870 test	CNN, 12-lead	AUC 0.93 for EF ≤35%
Attia 2019b [31]	Mayo Clinic sinus rhythm ECGs	180,922	CNN, 12-lead	AUC 0.87 (single ECG); 0.90 (multiple)
Raghunath 2021 [36]	Geisinger 12-lead ECGs	1.6 million ECGs	Deep neural network	AUC 0.85 for 1-year new-onset AF
Yao 2021, EAGLE [38]	120 US primary-care teams, cluster RCT	22,641	AI-ECG alert	New low EF 2.1% vs 1.6%; OR 1.32
Tsai 2025 [40]	Taiwanese inpatients, pragmatic RCT	13,631	AI-ECG alert	New low EF 1.5% vs 1.1%; HR 1.50
Perez 2019 [43]	Apple Watch users, US, single-arm	419,297	PPG irregular-rhythm algorithm	Notification PPV 0.84
Jabbour 2024 [37]	Montreal Heart Institute ECGs; MIMIC-IV external	145,323 patients	ECG deep learning + polygenic score + CHARGE-AF	5-year AF AUC 0.78 (ECG-AI); 0.59 (PGS)
Lin 2022 [44]	11 sites; SCOT-HEART prognostic cohort	921 training; 1,611 prognostic	Deep learning on CCTA	Plaque volume ICC 0.964 vs experts; MI HR 5.36
Poplin 2018 [25]	UK Biobank and EyePACS fundus images	284,335 training	Deep CNN	MACE AUC 0.70 vs 0.72 (SCORE)

5. Wearables and Consumer Devices

The smartwatch moves rhythm surveillance out of the clinic. An early study trained a deep neural network on smartwatch data from 9,750 participants; it detected atrial fibrillation with 98% sensitivity and 90% specificity in 51 patients undergoing cardioversion, but both fell to 68% in 1,617 ambulatory users [45]. The Apple Heart Study then tested the concept at scale. Over eight months it enrolled 419,297 participants, of whom 2,161 (0.52%) received an irregular pulse notification [43]. Among those who subsequently wore an ECG patch and returned an analysable recording, 34% had atrial fibrillation, and when an irregular tachogram coincided with patch monitoring the positive predictive value of a notification was 0.84 [43].

The study is often cited as proof that smartwatches work, but its design limits that inference. It had no control group and was not designed to measure outcomes such as stroke. Fewer than a quarter of notified participants returned an analysable patch, and only 5.9% of participants were aged 65 or older, the group in which atrial fibrillation is most common and in whom notifications were about six times more frequent [43]. The mSToPS trial, which tested a wearable ECG patch rather than a watch, found new atrial fibrillation diagnoses at one year in 6.3% of monitored participants against 2.4% of observational controls [46]. Whether consumer-detected atrial fibrillation, often brief and paroxysmal, carries the same stroke risk as clinically diagnosed disease remains uncertain.

Wearables are no longer limited to rhythm. As of September 2026 the FDA list of AI-enabled devices includes a hypertension notification feature for a consumer watch, cleared in September 2025, alongside digital stethoscope software and several ECG-based ejection fraction algorithms [47]. For India, where smartwatch ownership is likely to be concentrated in urban, higher-income groups, such devices are more likely to widen than to narrow existing gaps in detection unless public programmes adopt low-cost ECG hardware.

6. Imaging and Genomic Inputs

6.1 Retinal photographs

The retina is one of the few sites where small blood vessels can be photographed directly, and retinal vessel calibre carries cardiovascular information: in an individual-participant meta-analysis, narrower arterioles and wider venules were associated with higher coronary heart disease risk in women, though not significantly in men [48]. Deep learning first proved itself on fundus images in diabetic retinopathy screening [49], a topic developed in Chapter 10. Poplin and colleagues then trained models on 284,335 patients from UK Biobank and EyePACS to predict cardiovascular risk factors from fundus photographs alone [25]. The model estimated age with a mean absolute error of 3.26 years and systolic blood pressure with a mean absolute error of 11.23 mmHg, and it predicted smoking status with an AUC of 0.71 [25]. These

results are remarkable as biology, because neither age nor smoking is visible to an ophthalmologist on a fundus photograph.

As risk prediction, the results were more sober. For major adverse cardiac events over five years, the deep-learning model reached an AUC of 0.70, slightly below the 0.72 of the SCORE calculator, and the estimate rested on only 105 events in the validation set [25]. The model used attention maps to show which image regions drove each prediction, a form of saliency analysis [25]. Retinal imaging may yet prove useful where blood tests are hard to obtain, since a fundus camera in a primary-care or optometry setting could deliver a risk estimate without venepuncture. That use case requires validation against outcomes in South Asian populations, which had not been published in the sources reviewed for this chapter.

6.2 Coronary artery calcium and CT

Coronary artery calcium (CAC) scoring by non-contrast CT measures the calcified plaque burden directly. In intermediate-risk participants of the Multi-Ethnic Study of Atherosclerosis (MESA), CAC gave the largest improvement in risk classification among six novel markers, with a net reclassification improvement for coronary heart disease of 0.659, against 0.160 for family history, 0.102 for carotid intima-media thickness and 0.079 for C-reactive protein [50]. AI enters at two points. First, deep learning can automate calcium scoring, including opportunistic scoring on chest CT scans performed for other reasons, which could make CAC information available at no extra radiation dose. Second, on contrast-enhanced coronary CT angiography, algorithms can segment and quantify non-calcified plaque, estimate stenosis and derive a functional estimate of fractional flow reserve (CT-FFR) through computational modelling. In an international multicentre study, a deep-learning system quantified plaque and stenosis on coronary CT angiography with close agreement with expert readers (intraclass correlation 0.964 for total plaque volume) and with intravascular ultrasound (0.949), in about 5.7 seconds per patient against roughly 26 minutes for experts; in 1,611 SCOT-HEART participants, a deep-learning total plaque volume of 238.5 mm³ or more was associated with a hazard ratio of 5.36 for myocardial infarction after adjustment for stenosis and the ASSIGN score [44]. Coronary plaque analysis software, for example cvi42 Coronary Plaque (K251027), appears among the AI-enabled devices on the FDA list under the radiology panel [47].

Two cautions apply. CT is expensive and not widely available outside large Indian cities, so these tools are likely to influence referral and treatment decisions in tertiary centres rather than population screening. And measuring plaque more precisely is not the same as improving outcomes; that link has to be shown in its own right. Performance figures for individual commercial CT-FFR and plaque-analysis products are not reproduced here, and readers should consult the primary validation studies and the FDA decision summaries for each device.

6.3 Polygenic risk combined with machine learning

Polygenic risk scores add up the small effects of many common genetic variants to estimate inherited susceptibility to coronary artery disease. They have two appealing properties for prevention: they can be measured once, at any age, and they capture risk that is independent of the conventional factors. Machine learning offers a way to combine them with clinical, imaging and ECG data in a single model, but the gains so far are small. In 145,323 patients of the Montreal Heart Institute, an ECG deep-learning model predicted five-year incident atrial fibrillation with an AUC of 0.78; in a genotyped subgroup of 2,301 patients, the polygenic score alone gave an AUC of 0.59 and CHARGE-AF 0.62, and adding both to the ECG model improved fit while leaving the AUC at 0.76 to 0.77 [37]. The practical limitation for this chapter's readership is ancestry. Scores derived mainly from European genome-wide association studies tend to lose accuracy in other ancestries, and South Asian genetic architecture is underrepresented in the reference datasets. Until scores are calibrated in large Indian biobanks with outcome data, polygenic risk is best regarded as a research tool rather than a basis for treatment decisions in India.

7. Explainability, Calibration and Bias

7.1 Explaining predictions

A neural network that reads an ECG cannot state its reasoning. For tabular models, the most common remedy is feature attribution based on Shapley values from cooperative game theory, usually computed with the SHAP (Shapley additive explanations) framework [51]. Each variable receives a contribution to the individual prediction, and the contributions sum to the difference between that prediction and the average. For image and waveform models, gradient-based saliency methods such as Grad-CAM highlight the regions of the input that most influenced the output [52].

These tools are useful for model debugging and for checking that a model has not learned an artefact, such as a hospital-specific label or a scanner watermark. They are much less useful as a guarantee of safety at the bedside. A saliency map explains why a model produced one output for one tracing; it does not reveal the general rule the model has learned. Deep ECG models can also be manipulated: smoothed, almost imperceptible perturbations caused one network to classify 74% of atrial fibrillation test tracings as normal [53]. Explanations do not protect against such fragility. Prospective validation and post-market monitoring remain the more dependable safeguards.

7.2 Calibration: the neglected half of performance

Discrimination, measured by the AUC or C-statistic, asks whether people who have events receive higher scores than those who do not. Calibration asks whether a predicted risk of 10%

corresponds to an observed risk of 10%. For treatment thresholds, calibration matters more, because the decision to prescribe a statin depends on the absolute number. Most AI studies report the AUC prominently and calibration briefly, if at all. The failure of conventional scores in South Asians described in Section 2.4 is essentially a calibration failure: the ranking of individuals may be reasonable, but the absolute risks are too low [17].

The British experience with QRISK shows what sustained attention to calibration looks like. Independent external validation of QRISK2 in a separate primary-care database was carried out by researchers unconnected with its developers [28], and large electronic health record datasets allow models to be checked and recalibrated across regions, practices and subgroups [27]. AI models deserve the same scrutiny, and recalibration to a local population is often the simplest and most effective adjustment available.

7.3 Bias across ethnic groups

Algorithmic bias in health care became a mainstream concern after Obermeyer and colleagues showed that a commercial algorithm used to allocate care management in the United States systematically underestimated the needs of Black patients, because it predicted health-care costs rather than illness, and less money had historically been spent on Black patients with the same burden of disease [54]. The cardiovascular lesson is that the choice of outcome label can encode inequity before any algorithm is trained.

For AI-ECG, the question is whether a model trained mainly on patients of one ancestry performs equally well in others. Even where subgroup analyses within a US health system are reassuring, that reassurance does not automatically extend to Indian patients, whose ECG characteristics, body habitus and disease prevalence may differ. The SABRE findings for conventional scores, which underestimated risk most in South Asian women [19], are a reminder that subgroups defined by both sex and ancestry need separate reporting. The same caution applies to the rapid growth of AI-ECG in other regions: a model validated in Korea or the United States must be re-evaluated in India before use.

7.4 Dataset shift after deployment

Even a model that has been validated locally can deteriorate. ECG machines are replaced, filters and sampling rates change, laboratories switch assays, and the population tested shifts as referral patterns evolve. Treatment itself changes the outcome being predicted: if a risk model succeeds in directing statins to high-risk people, their observed event rate will fall, and the model will appear to overestimate risk in exactly the group it helped. These forms of drift are invisible unless performance is monitored continuously against outcomes. For conventional scores the remedy has been periodic recalibration, as the QRISK series and the regionally recalibrated SCORE2 and WHO charts illustrate [2,14,15]. For AI models the same discipline is needed, with the added complication that a retrained network may behave differently in ways

that are hard to anticipate. This is the problem that regulatory change control plans attempt to address [55].

Table 2.3: Challenges in AI-based cardiovascular risk prediction and practical mitigations

Challenge	Evidence	Mitigation
Poor calibration of imported scores in South Asians	PCE and QRISK3 underestimated the twofold excess risk in UK Biobank [17]; most Indian ACS patients scored low risk [21]	Recalibrate to local incidence; build Indian cohorts with adjudicated outcomes [5]
Unpredictable external performance	AUC fell from 0.93 to 0.82 in a Russian population, with sensitivity of 26.9% [41]; other external validations held up [33,42]	Validate in the intended population and setting; reset thresholds to local prevalence
Small gains from algorithm change alone	ML gains of 0.03-0.05 AUC, about half attributable to extra variables [22,23]	Prioritise new data types and local recalibration over algorithm complexity
Biased outcome labels	Cost-based label underestimated Black patients' needs [54]	Choose clinical outcome labels; audit performance by sex and ancestry subgroups [19]
Opaque deep models and adversarial fragility	Imperceptible ECG perturbations made 74% of AF tracings read as normal [53]	Use saliency for debugging only; rely on prospective trials and monitoring [52]
Prediction without action	About half of positive-alert patients in EAGLE did not receive an echocardiogram [38]	Design workflows and referral pathways alongside the model; report per CONSORT-AI [39]
Selected, unrepresentative cohorts	South Asians under 2% of UK Biobank [17]; only 5.9% of Apple Heart Study participants aged 65 or older [43]	Evaluate in community and LMIC populations; report demographic composition
Model change after approval	Learning systems drift over time	FDA predetermined change control plans; post-market surveillance [55]

8. Regulatory Status of Cardiovascular AI

The US Food and Drug Administration (FDA) maintains a public list of AI-enabled medical devices authorised for marketing. As of September 2026 the FDA states that it has authorised more than 1,600 such devices [56]. Cardiovascular devices form a substantial group, and the most recent entries show where the field is heading. Table 2.4 lists a selection of cardiovascular entries authorised between July 2025 and May 2026; functions are summarised from the device names and should be checked against the individual decision summaries.

Table 2.4: Selected AI-enabled cardiovascular devices on the FDA list, July 2025 to May 2026 [47]

Device	Manufacturer	Submission	Decision date	Function
ECG-AI Low Ejection Fraction	Anumana	K250652	28 Jul 2025	Low EF from 12-lead ECG
Hypertension Notification Feature	Apple	K250507	11 Sep 2025	Hypertension notification on a consumer watch
Bunkerhill ECG-EF	BunkerHill Health	K250649	19 Sep 2025	Low EF from ECG

Device	Manufacturer	Submission	Decision date	Function
DeepRhythmAI	Medicalgorithmics	K253141	11 Dec 2025	AI analysis of ECG rhythm recordings
eMurmur Heart AI	CSD Labs	K252284	19 Dec 2025	Heart murmur analysis
ECG-AI Pulmonary Hypertension	Anumana	K252360	28 Mar 2026	Pulmonary hypertension from ECG
ECG-AI Cardiac Amyloidosis	Anumana	K253801	7 Apr 2026	Cardiac amyloidosis from ECG
PCWP Analysis Software	Cardiosense	DEN250057	22 May 2026	Pulmonary capillary wedge pressure analysis (De Novo)

Three observations follow. First, all but one of these devices reached the market through the 510(k) pathway, which requires substantial equivalence to an existing device rather than evidence of improved outcomes. Second, the ECG now appears as a platform for multiple targets, including pulmonary hypertension and amyloidosis. This extends the line of work that began with ejection fraction [30]. Third, regulators have accepted that AI models may need updating after approval. The FDA's guidance on predetermined change control plans allows manufacturers to specify in advance the types of modification they will make and how those changes will be validated, without a new submission for each update [55]. Device authorisation in other jurisdictions, including India's Central Drugs Standard Control Organisation, falls outside the scope of this chapter, and global governance is discussed in Chapter 1.

9. Implementation in India

India's public system already has a framework into which risk prediction could fit. The national programme for non-communicable diseases previously known as NPCDCS was renamed the National Programme for Prevention and Control of Non-Communicable Diseases (NP-NCD) in May 2023. Under it, people aged 30 years and above are targeted for screening for diabetes, hypertension and common cancers through Ayushman Bharat Health and Wellness Centres, and the programme reported 724 district NCD clinics, 210 district cardiac care units and 6,110 community health centre NCD clinics in July 2023 [57]. The WHO HEARTS package offers simplified treatment protocols suited to team-based primary care [16]. In such a system, the immediate need is less for sophisticated AI than for a well-calibrated, low-cost risk estimate that frontline workers can apply. The non-laboratory WHO chart, recalibrated for South Asia, fits that description [2].

Where, then, might AI help? Four uses appear realistic in the near term. The first is better calibration: machine learning methods can be used to recalibrate and update risk equations as Indian cohorts accumulate outcome data, a direct response to the problem shown by Gupta and colleagues [5]. The second is opportunistic screening from existing tests. ECGs are already widely performed in Indian primary and secondary care, and an AI-ECG that flags left ventricular dysfunction or hidden atrial fibrillation could turn a routine test into a screening

tool at almost no marginal cost, provided that referral to echocardiography is available [38,40]. The third is task-sharing: automated interpretation of ECGs or fundus images could support community health workers and medical officers in centres without a specialist. The fourth is follow-up: risk models can be used to prioritise people for recall when resources are scarce.

Each use comes with conditions. Much of India's ECG output is still printed on thermal paper rather than stored as digital waveforms, so AI-ECG at scale would require either digital acquisition or validated methods for converting paper or photographed tracings into signals, and the latter introduce errors of their own. Echocardiography is concentrated in district and tertiary hospitals, so a positive screen in a primary health centre is useful only if a referral pathway exists and the patient can travel. Imported models must be validated on Indian data before deployment, because the external validation experience in Section 4.4 shows that performance can fall unexpectedly [41]. Outcome labels must reflect disease rather than access to care, to avoid reproducing the bias documented by Obermeyer and colleagues [54]. Data infrastructure, interoperability and federated approaches to model training are addressed in Chapters 8 and 10. Finally, any AI tool should be introduced, where possible, through a pragmatic evaluation such as a cluster-randomised or stepped-wedge design, following the example of EAGLE [38].

10. Conclusion and Future Directions

The past decade has clarified what AI can and cannot do for cardiovascular risk prediction. Applied to the variables that clinicians already collect, machine learning has improved discrimination modestly and consistently, from about 0.72 to 0.76-0.77 in large British datasets, and much of that gain comes from richer data rather than from the algorithm [22,23]. Applied to new signals, AI has opened genuinely new possibilities. The AI-ECG for low ejection fraction is the strongest example: it was developed on large paired datasets, validated externally with mixed results, and tested in two pragmatic randomised trials that increased diagnostic yield [30,38,40,41]. Retinal and CT-based models are promising but have not yet shown clear superiority over established scores for hard outcomes [25].

For India and other low- and middle-income countries, the most pressing problem lies upstream of AI. Conventional scores underestimate risk in South Asians, local cohorts with adjudicated outcomes are scarce, and recalibrations disagree with one another [5,17,21]. Several priorities follow. Prospective Indian cohorts with linked ECG, imaging and outcome data are needed to derive and recalibrate models. AI studies should report calibration and subgroup performance alongside discrimination, and they should follow reporting standards for trials of AI interventions [39]. Opportunistic AI-ECG screening deserves pragmatic evaluation within NP-NCD, linked to accessible echocardiography and treatment. Multimodal models that combine clinical, ECG, imaging and genomic inputs should be tested against simple recalibrated scores

rather than against outdated comparators. And regulatory oversight must keep pace with models that change after approval [55].

AI will not replace the risk score. The more probable future is a layered approach in which a simple, well-calibrated score serves the whole population and AI-derived signals refine risk where data and services allow. Whether that combination improves cardiovascular outcomes in Indian patients is an empirical question, and it should be answered with trials rather than assumed.

REFERENCES

1. Global, regional, and national burden of cardiovascular diseases and risk factors in 204 countries and territories, 1990-2023. J Am Coll Cardiol. 2025;86(22):2167-243. https://doi.org/10.1016/j.jacc.2025.08.015
2. WHO CVD Risk Chart Working Group. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. Lancet Glob Health. 2019;7(10):e1332-45. https://doi.org/10.1016/S2214-109X(19)30318-3
3. India State-Level Disease Burden Initiative Collaborators. Nations within a nation: variations in epidemiological transition across the states of India, 1990-2016 in the Global Burden of Disease Study. Lancet. 2017;390(10111):2437-60. https://doi.org/10.1016/S0140-6736(17)32804-0
4. Patel V, Chatterji S, Chisholm D, Ebrahim S, Gopalakrishna G, Mathers C, et al. Chronic diseases and injuries in India. Lancet. 2011;377(9763):413-28. https://doi.org/10.1016/S0140-6736(10)61188-9
5. Gupta P, Prieto-Merino D, Ajay VS, Singh K, Roy A, Krishnan A, et al. Cardiovascular risk prediction in India: comparison of the original and recalibrated Framingham prognostic models in urban populations. Wellcome Open Res. 2019;4:71. https://doi.org/10.12688/wellcomeopenres.15137.2
6. Joshi P, Islam S, Pais P, Reddy S, Dorairaj P, Kazmi K, et al. Risk factors for early myocardial infarction in South Asians compared with individuals in other countries. JAMA. 2007;297(3):286-94. https://doi.org/10.1001/jama.297.3.286
7. Xavier D, Pais P, Devereaux PJ, Xie C, Prabhakaran D, Reddy KS, et al. Treatment and outcomes of acute coronary syndromes in India (CREATE): a prospective analysis of registry data. Lancet. 2008;371(9622):1435-42. https://doi.org/10.1016/S0140-6736(08)60623-6
8. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. https://doi.org/10.1038/s41591-018-0300-7
9. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837-47. https://doi.org/10.1161/01.cir.97.18.1837
10. D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743-53. https://doi.org/10.1161/CIRCULATIONAHA.107.699579

11. Goff DC Jr, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014;129(25 Suppl 2):S49-73. https://doi.org/10.1161/01.cir.0000437741.48606.98
12. Khan SS, Matsushita K, Sang Y, Ballew SH, Grams ME, Surapaneni A, et al. Development and validation of the American Heart Association's PREVENT equations. Circulation. 2024;149(6):430-49. https://doi.org/10.1161/CIRCULATIONAHA.123.067626
13. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, De Backer G, De Bacquer D, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24(11):987-1003. https://doi.org/10.1016/S0195-668X(03)00114-3
14. SCORE2 working group and ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. Eur Heart J. 2021;42(25):2439-54. https://doi.org/10.1093/eurheartj/ehab309
15. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357:j2099. https://doi.org/10.1136/bmj.j2099
16. World Health Organization. HEARTS: technical package for cardiovascular disease management in primary health care: evidence-based treatment protocols. Geneva: World Health Organization; 2018 (WHO/NMH/NVI/18.2). Available from: https://iris.who.int/bitstream/handle/10665/260421/WHO-NMH-NVI-18.2-eng.pdf
17. Patel AP, Wang M, Kartoun U, Ng K, Khera AV. Quantifying and understanding the higher risk of atherosclerotic cardiovascular disease among South Asian individuals: results from the UK Biobank prospective cohort study. Circulation. 2021;144:410-22.
18. Bhopal R, Fischbacher C, Vartiainen E, Unwin N, White M, Alberti G. Predicted and observed cardiovascular disease in South Asians: application of FINRISK, Framingham and SCORE models to Newcastle Heart Project data. J Public Health (Oxf). 2005;27(1):93-100. https://doi.org/10.1093/pubmed/fdh202
19. Tillin T, Hughes AD, Whincup P, Mayet J, Sattar N, McKeigue PM, et al. Ethnicity and prediction of cardiovascular disease: performance of QRISK2 and Framingham scores in a U.K. tri-ethnic prospective cohort study (SABRE). Heart. 2014;100(1):60-7. https://doi.org/10.1136/heartjnl-2013-304474
20. Chow CK, Joshi R, Celermajer DS, Patel A, Neal BC. Recalibration of a Framingham risk equation for a rural population in India. J Epidemiol Community Health. 2009;63(5):379-85. https://doi.org/10.1136/jech.2008.077057
21. Das S, Biswas L, Mukherjee T, Biswas K, Pal SK. Predictive ability of Framingham risk score in Indian population: a retrospective study in a tertiary care hospital in patients with first acute coronary syndrome. Asian J Med Sci. 2022;13(8):133-9. https://doi.org/10.3126/ajms.v13i8.45309
22. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944. https://doi.org/10.1371/journal.pone.0174944

23. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PLoS One. 2019;14:e0213653. https://doi.org/10.1371/journal.pone.0213653
24. Hill NR, Ayoubkhani D, McEwan P, Sugrue DM, Farooqui U, Lister S, et al. Predicting atrial fibrillation in primary care using machine learning. PLoS One. 2019;14(11):e0224582. https://doi.org/10.1371/journal.pone.0224582
25. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2(3):158-64. https://doi.org/10.1038/s41551-018-0195-0
26. Christopoulos G, Graff-Radford J, Lopez CL, Yao X, Attia ZI, Rabinstein AA, et al. Artificial intelligence-electrocardiography to predict incident atrial fibrillation: a population-based study. Circ Arrhythm Electrophysiol. 2020;13(12):e009355. https://doi.org/10.1161/CIRCEP.120.009355
27. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140. https://doi.org/10.1136/bmj.i3140
28. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. BMJ. 2010;340:c2442. https://doi.org/10.1136/bmj.c2442
29. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65-9. https://doi.org/10.1038/s41591-018-0268-3
30. 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(1):70-4. https://doi.org/10.1038/s41591-018-0240-2
31. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, 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. 2019;394(10201):861-7. https://doi.org/10.1016/S0140-6736(19)31721-0
32. Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G, et al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. Circ Arrhythm Electrophysiol. 2019;12(9):e007284. https://doi.org/10.1161/CIRCEP.119.007284
33. Kwon JM, Lee SY, Jeon KH, Lee Y, Kim KH, Park J, et al. Deep learning-based algorithm for detecting aortic stenosis using electrocardiography. J Am Heart Assoc. 2020;9(7):e014717. https://doi.org/10.1161/JAHA.119.014717
34. Kwon JM, Kim KH, Jeon KH, et al. Development and validation of deep-learning algorithm for electrocardiography-based heart failure identification. Korean Circ J. 2019;49(7):629-39. https://doi.org/10.4070/kcj.2018.0446
35. Galloway CD, Valys AV, Shreibati JB, Treiman DL, Petterson FL, Gundotra VP, et al. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. JAMA Cardiol. 2019;4(5):428-36. https://doi.org/10.1001/jamacardio.2019.0640

36. Raghunath S, Pfeifer JM, Ulloa-Cerna AE, Nemani A, Carbonati T, Jing L, et al. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation-related stroke. Circulation. 2021;143(13):1287-98.

37. Jabbour G, Nolin-Lapalme A, Tastet O, Corbin D, Jordà P, Sowa A, et al. Prediction of incident atrial fibrillation using deep learning, clinical models, and polygenic scores. Eur Heart J. 2024;45(46):4920. https://doi.org/10.1093/eurheartj/ehae595

38. 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. 2021;27:815-9. https://doi.org/10.1038/s41591-021-01335-4

39. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-74. https://doi.org/10.1038/s41591-020-1034-x

40. Tsai DJ, Lin C, Liu WT, Lee CC, Chang CH, Lin WY, et al. Artificial intelligence-assisted diagnosis and prognostication in low ejection fraction using electrocardiograms in inpatient department: a pragmatic randomized controlled trial. BMC Med. 2025;23:342. https://doi.org/10.1186/s12916-025-04190-z

41. 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;329:130-5. https://doi.org/10.1016/j.ijcard.2020.12.065

42. König S, Hohenstein S, Nitsche A, Pellissier V, Leiner J, Stellmacher L, et al. Artificial intelligence-based identification of left ventricular systolic dysfunction from 12-lead electrocardiograms: external validation and advanced application of an existing model. Eur Heart J Digit Health. 2024;5(2):144-51. https://doi.org/10.1093/ehjdh/ztad081

43. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med. 2019;381(20):1909-17. https://doi.org/10.1056/NEJMoa1901183

44. Lin A, Manral N, McElhinney P, Killekar A, Matsumoto H, Kwiecinski J, et al. Deep learning-enabled coronary CT angiography for plaque and stenosis quantification and cardiac risk prediction: an international multicentre study. Lancet Digit Health. 2022;4(4):e256-65.

45. Tison GH, Sanchez JM, Ballinger B, Singh A, Olgin JE, Pletcher MJ, et al. Passive detection of atrial fibrillation using a commercially available smartwatch. JAMA Cardiol. 2018;3:409-16. https://doi.org/10.1001/jamacardio.2018.0136

46. Steinhubl SR, Waalen J, Edwards AM, et al. Effect of a home-based wearable continuous ECG monitoring patch on detection of undiagnosed atrial fibrillation: the mSToPS randomized clinical trial. JAMA. 2018;320(2):146-55.

47. U.S. Food and Drug Administration. List of artificial intelligence-enabled medical devices [Internet]. Silver Spring (MD): FDA; 2026 [cited 2026 Nov 3]. Available from: https://www.fda.gov/medical-devices/artificial-intelligence-enabled-medical-devices/list-artificial-intelligence-enabled-medical-devices

48. McGeechan K, Liew G, Macaskill P, Irwig L, Klein R, Klein BEK, et al. Meta-analysis: retinal vessel caliber and risk for coronary heart disease. Ann Intern Med. 2009;151(6):404-13. https://doi.org/10.7326/0003-4819-151-6-200909150-00005

49. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402-10. https://doi.org/10.1001/jama.2016.17216

50. Yeboah J, et al. Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individuals. JAMA. 2012;308:788-95.

51. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017. Available from: https://papers.nips.cc/paper/7062-a

52. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017. p. 618-26. https://doi.org/10.1109/ICCV.2017.74

53. Han X, Hu Y, Foschini L, Chinitz L, Jankelson L, Ranganath R. Deep learning models for electrocardiograms are susceptible to adversarial attack. Nat Med. 2020;26(3):360-3. https://doi.org/10.1038/s41591-020-0791-x

54. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53. https://doi.org/10.1126/science.aax2342

55. U.S. Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: guidance for industry and Food and Drug Administration staff [Internet]. Silver Spring (MD): FDA; 2024 [updated 2025 Aug 18; cited 2026 Nov 3]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence

56. U.S. Food and Drug Administration. Artificial intelligence-enabled medical devices [Internet]. Silver Spring (MD): FDA; 2026 [updated 2026 Sep 22; cited 2026 Nov 3]. Available from: https://www.fda.gov/medical-devices/digital-health-center-excellence/artificial-intelligence-enabled-medical-devices

57. Ministry of Health and Family Welfare, Government of India. Lok Sabha Unstarred Question No. 272: answered on 21 July 2023 [Internet]. New Delhi: Lok Sabha Secretariat; 2023 [cited 2026 Nov 3]. Available from: https://eparlib.nic.in/bitstream/123456789/2503823/1/AU272.pdf