Mathematical and Computational Applications

MDPI

Supplementary Information

Electrocardiogram Foundation Models: From Representation Learning to Clinical Evidence

Abbreviations used in this Supplementary Information: AF, atrial fibrillation; AI, artificial intelligence; AI-ECG, AI-enabled electrocardiography; AUC, area under the curve; AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; CLD, chronic liver disease; CT, computed tomography; ECG, electrocardiogram; ED, emergency department; EF, ejection fraction; EHR, electronic health record; FM, foundation model; FPR, false-positive rate; HR, hazard ratio; ICER, incremental cost-effectiveness ratio; IQR, interquartile range; IRR, incidence rate ratio; LoRA, low-rank adaptation; LV, left ventricular; LVEF, left ventricular ejection fraction; LVSD, left ventricular systolic dysfunction; NPV, negative predictive value; OR, odds ratio; PPV, positive predictive value; QALY, quality-adjusted life-year; RCT, randomized controlled trial; RR, risk ratio; STEMI, ST-elevation myocardial infarction; TPR, true-positive rate.

Supplementary Note 1. Literature identification, source verification and registry search

Design. This Review is a critical, non-systematic synthesis. We identified sources by targeted searches of the web, preprint servers (arXiv, medRxiv), publisher sites, PubMed-derived records and bibliographic databases (Crossref, OpenAlex, DataCite, Europe PMC where accessible) up to 21 October 2026, supplemented by citation tracking from primary reports and recent reviews [1-4]. We did not register a protocol, apply formal eligibility criteria or perform dual screening, so the evidence base may be incomplete and is subject to selection by the authors.

Scope and limits of the search. The searches were targeted, not run with one prespecified search string, and the query terms used for the literature search were not fixed in advance; the registry queries are listed in Supplementary Table S12. The sources sought, stated retrospectively because no eligibility criteria were prespecified, were reports of models presented as ECG foundation models or as reusable pretrained ECG representations; evaluations of such models by their developers or by others; validation studies of AI-enabled ECG relevant to generalization, calibration or reliability; and randomized or prospective evaluations of AI-enabled ECG in clinical care. Sources were retained when they bore on one of the evidence questions of the Review; we did not record excluded records and therefore report no screening counts. Statements in the Review that evidence was not found, not identified or not reported refer only to the sources examined within this search. A further targeted search made during revision is reported separately in Supplementary Note 4 and is not part of the search described here.

Verification. For each quantitative claim we recorded the depth at which the source was read: the full text of a journal article or preprint, the journal page (abstract, figure legends), the abstract only (from the publisher, an institutional repository or a bibliographic database), or a secondary report. Claims that rest on abstracts or secondary

reports are labelled in the text and tables, and evidence-matrix cells that rest on them are coded as unverified (Supplementary Table S6). Where a source reported inconsistent values (for example between an abstract and the Results), we report both. Source verification included the full-text reports of DeepECG, EAGLE, DULCE, ARISE, and the inpatient, mortality-alert, potassium-alert, and SPEC-AI trials; the ECGFounder author manuscript; the published corrections to DULCE and DeepECG; the PubMed abstract of ECG-CLIP; and the operative text of Regulation (EU) 2026/1744. Values were taken as published and were not recomputed, except where labelled as our arithmetic. Bibliographic metadata were checked against Crossref, OpenAlex, DataCite, publisher pages or regulatory websites, and the verification status of each record is stored in the master RIS library (keyword "Status" and note field).

Evaluator independence. Independence was assessed separately for each model–evaluation pair from author lists, affiliations, funding and competing-interests statements in the texts read (Supplementary Tables S7 and S10). Competing scientific or financial interests were recorded and disclosed but were not by themselves disqualifying. Financial relationships not stated in those texts could not be assessed.

Registry search. On 21 October 2026 we searched ClinicalTrials.gov (API v2, all fields) and ISRCTN (Supplementary Table S12) for foundation-model terms, self-supervised and pretrained ECG models, named ECG foundation models and large language models applied to the ECG. Relevant interventional records were classified from their registered intervention descriptions into trials that may evaluate an ECG foundation model directly, trials of conventional or unspecified algorithms from a foundation-model developer group, and trials of general-purpose multimodal language models. WHO ICTRP, ChiCTR, EU CTIS and other registries were not searched. Statements about the absence of trials are limited to these two registries.

The evidence matrix was developed using the coding criteria in Supplementary Tables S6–S8 and the intended-use definitions for five clinical applications in Supplementary Table S9. The initial extraction and organization of supporting evidence and preparation of provisional classifications were conducted by the authors. Generative AI was subsequently used to cross-check the extracted information and identify potential inconsistencies. The first rater manually reviewed all 116 evidence cells (Supplementary Table S5), verified the supporting sources, resolved discrepancies, and finalized the classifications. Verification was based on original publications and supplementary materials or, where full texts were unavailable, abstracts and other documented sources. Findings insufficiently supported by the available evidence were classified as unverified, while the absence of identified evidence was distinguished from demonstrated failure. Supplementary Table S11 documents the supporting sources and rationale for cells other than those classified solely as “not identified.” This paragraph describes the first rater’s coding. A second author independently coded all 116 cells during revision, followed by agreement analysis and adjudication of disagreements, as reported in Supplementary Note 5. Table S5 presents the consensus codes. The matrix remains an illustrative appraisal, not a validated assessment instrument.

Preparation for independent assessment (plan written before the second coding and retained unchanged). Supplementary Data 1 contains a blank coding form for all 116 cells (21 backbone and 95 use cells), without our codes, together with the coding guide and source list. Raters code all 116 cells; the 26 rater-training cells (7 for the ECG-FM backbone and 19 for the structural heart disease use) are excluded from the agreement analysis by a design fixed before any rating, so that the analysis uses 90 cells; descriptive codes are reported for all 116 and both denominators are to be reported. Training cells are so named to avoid confusion with statistical calibration of model risk. We propose that two raters with clinical-epidemiology or machine-learning training, not involved in this Review,

code every cell independently from the primary sources, blind to our codes and to each other, after training on one backbone and one use that are then excluded from analysis. Agreement would be reported separately for the evidence state (V, U, N, X) and the evaluator code, as observed agreement, Cohen's kappa and Gwet's AC1 (because N dominates, kappa alone is unstable) with bootstrap 95% confidence intervals over cells, and as agreement on each requirement across columns. Because cells share models, uses and sources, resampling cells as if independent may give intervals that are too narrow; a cluster bootstrap over model–use columns, or descriptive statistics with justified intervals, is reported as a sensitivity analysis. Disagreements would be resolved by a third rater, and any requirement with poor agreement would be redefined before reuse of the guide. The analysis plan is ours and has not been registered. The procedure actually followed differed from this plan, as set out in Supplementary Note 5.

Supplementary Note 2. Methodological notes on comparative studies

ICLR 2026 benchmark [5]. The abstract and introduction describe 26 clinical tasks, whereas the main results table displays 24 result rows in seven categories, and the paper does not define its unit of counting. Two displayed rows each combine three PTB-XL label sets, and several rows bundle many endpoints (for example 11 echocardiographic findings, 158 cardiac discharge diagnoses and 918 non-cardiac discharge diagnoses). Counting every PTB-XL label set separately gives 28 rows, and counting only the first PTB-XL row's three label sets separately gives 26. We therefore describe the benchmark by its 24 displayed rows. The benchmark used native input formats (for example S4 at 100 Hz and ECG-CPC at 240 Hz) and one training recipe (AdamW, learning rate 10^{-3}, 100 epochs, layer-dependent learning rates) without model-specific tuning. Its authors developed ECG-CPC and the S4 baseline. On PTB-XL with all labels, AUROC was 0.949 for ECG-CPC, 0.941 for S4, 0.940 for ECG-JEPA, 0.934 for ECGFounder, 0.929 for a supervised convolutional baseline, 0.927 for ECG-FM, 0.925 for MERL, 0.915 for HuBERT-ECG, 0.908 for ST-MEM and 0.889 for ECGFM-KED. In the label-efficiency experiment, learning curves of 1 - macro AUROC were fitted with a power law plus a residual term. The 3.3- to 9-fold gains stated in the abstract correspond, by our calculation, to the inverses of the smallest ratio for each model (1/0.30 for ECGFounder, 1/0.21 for ECG-CPC and 1/0.11 for ECG-JEPA); the derivation is not shown in the paper.

DeepECG [6, 7]. The self-supervised model was pretrained on about 1.9 million ECGs, including CODE-15% and the MIMIC-IV training partition, whereas the supervised model was trained on about 1.02 million ECGs from one institution, with a 670-run Bayesian hyperparameter sweep whose counterpart for the self-supervised model is not described. The published Methods (full text read, 21 October 2026) state that development-hospital ECGs were split randomly 70/10/20 at patient level, that MIMIC-IV was split 70/30 with patient partitioning, that per-label thresholds were optimized on the development validation split and applied unchanged to all cohorts, and that calibration was assessed with a weighted Brier score. We also examined the supplementary tables reporting dataset-specific results and weighted Brier scores. However, group-level fairness results were presented only in a figure that was unavailable for our assessment. The article names seven private health systems and four public repositories. For reduced ejection fraction it reports AUROC 0.926 versus 0.917 in the abstract and Results and 0.928 versus 0.900 in its structured summary, which matches the preprint abstract; the supplementary table of internal results confirms 0.926 versus 0.917, and the structured-summary values do not appear in the tables we read. The supplementary tables also show different operating points for the two models (internal overall sensitivity 0.994 and specificity 0.795 for the self-supervised model, against 0.959 and 0.952) and do not state which of two listed thresholds produced them. Dataset counts differ between parts of the article (11 and 13 datasets), and

the long QT genotype analysis is given as 127 ECGs in the abstract and 334 in the Results. The authors state that P values were not adjusted for multiple comparisons. The article also benchmarks ECGFounder and ECG-FM on overlapping labels and reports models trained from random initialization. Its main-text long QT comparison of pretrained with randomly initialized models (0.875 versus 0.602) is labelled in Supplementary Table S37 of the DeepECG article as a comparison between two randomly initialized models, so we do not use it; for ejection fraction below 50% the randomly initialized supervised model exceeded both fine-tuned models (0.909 against 0.885 and 0.887), but not for 40% or below.

ECG-Fyler [8]. When ECGFounder was adapted with 100 labelled paediatric examples, its internal AUROC for reduced ejection fraction was 0.47 (95% CI 0.46–0.49), against 0.50 for a ResNet trained from scratch, 0.63 for a masked autoencoder pretrained on the paediatric data and 0.80 for ECG-Fyler. With 100 adult examples, ECGFounder reached 0.72. The paper does not state whether encoders were frozen or how comparator hyperparameters were set. Candidate explanations, none tested, are a mismatched adaptation recipe, mismatched input handling (sampling rate, lead order, length, amplitude scaling), instability with 100 examples, and absence of paediatric information in an adult-trained representation.

ECG-JEPA [9]. Its developers retrained some baselines on their own pretraining data, took others from published results and used a released ST-MEM checkpoint, and pretraining compute was not matched (100 epochs for ECG-JEPA against 300–800 for several baselines).

HuBERT-ECG [10]. Several comparator values were taken from other publications with different label scopes. The smallest model outperformed the larger ones on some tasks (for example, AUROC 0.894 against 0.756 and 0.774 on CPSC-Extra).

Frozen-embedding benchmark [11]. In the smallest training-set condition on CODE-15%, median F1 for atrial fibrillation was 0.97 for ECGFounder, 0.94 for ECG-FM, 0.84 for HuBERT-ECG small and 0.67 for ECG-JEPA. Some models had been pretrained on parts of the evaluation data.

Supplementary Note 3. Study-level notes on generalization evidence

Record-level separation. A single-author biometric preprint reported closed-set identification of individuals with Rank@1 0.95 internally and 0.83 and 0.69 in MIMIC-IV-ECG and HEEDB cohorts. Identification fell over a one-to-five-year gap, from 0.785 to 0.643 and from 0.686 to 0.556 [12]. We cite it only to show that individual identity is recoverable across institutions; we have not read how its partitions were constructed. The published ECGFounder article describes the Harvard–Emory database as 10,771,552 ECGs from 1,818,247 subjects, the whole database, and reports, after excluding unreadable, missing and unmatched records, 7,519,035 ECGs from 1,319,128 patients for development and 834,926 from 146,570 for testing; its arXiv version gave 1,725,643 and 265,818 patients, which together exceed the database total and so imply overlap or an error in that version. Neither version describes how the split was made, and the fact that the published subtotals fit under the total does not establish a patient-level split. A published correction (September 2025) revised the description of the loss function for missing positive labels and replaced the table of PTB-XL results; we have not re-derived the average PTB-XL AUROC from the corrected table [13]. A Korean left ventricular dysfunction model split patients randomly 8:1:1 [14].

Pretraining–evaluation overlap. In the largest independent benchmark, HuBERT-ECG's pretraining included PTB, CPSC2018 and PTB-XL; ECG-FM's included CPSC2018, PTB-XL and MIMIC-IV-ECG; and MERL and ECGFM-KED were pretrained on MIMIC-IV-ECG, all of which were benchmark datasets, with no overlap check reported [5]. In

DeepECG, a 30% held-out MIMIC-IV partition, separated by patient, was among the external public test sets while the training partition contributed to pretraining [6].

Hyperkalaemia [15]. The endpoint was potassium above 5.5 mmol/L, with each ECG paired to the nearest non-haemolysed potassium result within one hour. The mean intervals between ECG and potassium measurement were 28, 24 and 22 minutes in the internal, temporal and external cohorts. Negative predictive value in the external cohort was above 99.3%. The temporal sensitivity was 85.05% (95% CI 81.84–87.77%), and external sensitivity, specificity and PPV at the primary threshold are not reported.

Korean left ventricular dysfunction model [14]. AUROC was 0.953 internally and 0.947 externally for 12-lead ECGs. In follow-up pairs it was 0.945 and 0.910 before adjustment with earlier echocardiographic results, and 0.956 and 0.940 after. Specificity at the Youden threshold set on development data was 0.852 internally and 0.763 externally before adjustment, and 0.912 and 0.859 after.

ECG-FM [16]. The UHN-ECG cohort comprised 622,000 ECGs from 211,000 patients (2010–2018). Of the recordings, 88.8% were at 500 Hz and the remainder at 250 Hz, and 12.8% were labelled as of poor quality. Per-task results are in supplementary tables that we did not see.

Monachino et al. [17, 18]. The study compared partial fine-tuning, BitFit, LayerNorm tuning and linear probing with full fine-tuning of the Self-DANA backbone. We read the preprint abstract and the supplementary information. According to the supplementary information, the out-of-domain sets were small public datasets. We found no report of the number of random seeds or of the tuning budget.

Test–retest [19]. Instability, defined as a difference of more than 0.5 in the ensemble score, was more likely in patients younger than 65 years (OR 2.52, 95% CI 2.38–2.67) and in inpatients (OR 1.91), slightly more likely in men (OR 1.10), and only weakly associated with race and ethnicity. Cohen's kappa for binary outputs was 0.75–0.82 for the ensemble and 0.57–0.72 for individual networks. For re-rendered images, Spearman correlation was 0.969–0.998; for real-world captures it was 0.942–0.962.

Demographic transportability [20, 21]. The foundation model for left ventricular systolic dysfunction was pretrained on 983,200 ECGs and fine-tuned on 44,815 paired ECG–echocardiogram records, with a test cohort of 8,963 patients. In the earlier task-specific study, the derivation cohort was 96.2% non-Hispanic White, and a separate network identified race from the ECG with AUCs of 0.84 for Black and 0.76 for non-Hispanic White patients in a five-class problem.

TolerantECG [22]. This developer preprint reports first or second rank across signal conditions on PTB-XL and first on the MIT-BIH arrhythmia database. The abstract gives no metric values and does not name its baselines.

Table S1. Comparative evidence: study-level extraction.

Study [ID]; evaluator; status	Models compared	Pretraining data	Task(s); family	Evaluation data; separation	Adaptation; tuning	Key results (95% CI where reported)	Comparability limitations
A. Multi-model comparisons by independent or partly independent evaluators							
Al-Masud et al., 2026 [5]; I, with disclosed interest (authors developed ECG-CPC, a compared model; no competing-interests statement); peer-reviewed conference	8 FMs: ECGFounder, ECG-JEPA, ST-MEM, MERL, ECGFM-KED, HuBERT-ECG, ECG-FM, ECG-CPC. Supervised from scratch: S4 (2.2M parameters), Net1D (33.8M)	As released (e.g., HEEDB 10.7M ECGs for ECGFounder; MIMIC-IV-ECG ~800k for MERL); ECG-CPC 3.8M parameters, HEEDB	24 result rows (26 tasks stated; counting unit not defined; two rows each combine 3 PTB-XL label sets; several rows bundle many endpoints), 7 categories: adult and paediatric interpretation, structure/function, cardiac and non-cardiac outcomes, acute care, patient characteristics; F1-F5	12 datasets incl. PTB-XL, CODE-15%, ZZU pECG, EchoNext, MIMIC-IV-ECG; in-distribution test splits only. Several benchmark datasets were in pretraining corpora (HuBERT-ECG, ECG-FM, MERL, ECGFM-KED); no overlap audit	Fine-tuning, frozen and linear probe. One recipe for all models (AdamW, lr 1e-3, 100 epochs, batch 64, layer-wise lr); no model-specific tuning reported; native input formats (S4 100 Hz, ECG-CPC 240 Hz)	PTB-XL (all) AUROC: ECG-CPC 0.949; S4 0.941; ECG-JEPA 0.940; ECGFounder 0.934; Net1D 0.929; ECG-FM 0.927; MERL 0.925; HuBERT-ECG 0.915; ST-MEM 0.908; ECGFM-KED 0.889. Acute care: no FM significantly > S4. Label efficiency (one experiment: EchoNext, 11 echo findings, metric 1 - macro AUROC, lower is better; N = 250–1,000): labels needed to match S4 trained on N were 0.30–0.62 N (ECGFounder), 0.11–0.42 N (ECG-JEPA), 0.21–0.40 N (ECG-CPC); reported as 3.3–9-fold; all three FMs had lower error than S4 at every training size. Pairwise differences by bootstrap (1,000)	Best model is the evaluators' own; single recipe may favour some architectures; in-distribution only; stated 26 tasks not uniquely reconstructable from 24 displayed rows; label-efficiency estimate limited to 3 FMs, 1 dataset, 1 endpoint set and N = 250–1,000; CIs for individual AUROCs not given in main table
Filice et al., 2026 [11]; I; preprint	ECGFounder, ECG-FM, ECG-JEPA, HuBERT-ECG small/base/large. No supervised baseline	As released	Conduction disturbance (CD), atrial fibrillation (AF); F1	PTB-XL, CODE-15%, Georgia, Chapman (public); overlap with some models' pretraining corpora not excluded	Frozen embeddings + classical classifiers; training-set sizes from XS to L	Median F1, XS/L. PTB-XL CD: ECGFounder 0.81/0.87; ECG-FM 0.75/0.83; ECG-JEPA 0.76/0.80; HuBERT S/B/L 0.75/0.79, 0.73/0.77, 0.74/0.78. CODE-15% AF: ECGFounder 0.97/0.96; ECG-FM 0.94/0.96; ECG-JEPA 0.67/0.77; HuBERT S/B/L 0.84/0.89, 0.77/0.86, 0.82/0.88. Clustering of embeddings by source dataset (ARI, CD/AF): ECG-JEPA 0.98/1.00; HuBERT 0.42–0.47/0.66–0.71; ECG-FM 0.31/0.20; ECGFounder 0.03/0.12	Frozen regime only; no supervised comparator; rhythm and conduction tasks only, which match ECGFounder's supervised labels; preprint

Study [ID]; evaluator; status	Models compared	Pretraining data	Task(s); family	Evaluation data; separation	Adaptation; tuning	Key results (95% CI where reported)	Comparability limitations
Taleshinosrati et al., 2026 (FOUND-AF) [23]; I; preprint	ECGFounder, ECG-JEPA and other released ECG FMs. No supervised baseline under the same protocol	As released	Binary AF detection; F1	AFDB, CinC2017, CPSC2021, LTAFDB (public; different devices and sampling rates)	Frozen encoders + gradient-boosted trees (XGBoost)	ECGFounder first by F1 on all 4 datasets: 97.87, 92.77, 99.50, 95.68%; AUROC 0.9968, 0.9946, 0.9999, 0.9804. Significant vs all comparators except ECG-JEPA on AFDB	Near-ceiling binary task; specialized models in the literature similar or higher on some datasets; bibliographic metadata only partly verified (index lookup rate-limited); preprint
Do et al., 2026 [24]; I (developers of the adaptation method, not of the FMs); preprint, accepted to Canadian AI 2026	ECG-FM with in-domain self-supervised adaptation (frozen, partial, full fine-tuning, LoRA); ECGFounder; HuBERT-ECG; ResNet1D from scratch; XGBoost on features	ECG-FM as released + continued self-supervised pretraining on EchoNext	Multi-label structural heart disease defined by echocardiography; F2	EchoNext (public, single site), standard split; no external validation	Several adaptation strategies; models not matched for size	Macro AUROC/AUPRC: XGBoost 0.8017/0.3289; ResNet1D 0.8246/0.3833; adapted ECG-FM frozen 0.8233/0.3670; partial FT 0.8509/0.4261; same without in-domain adaptation 0.8387/0.4048; full FT 0.8484 (AUROC); LoRA 0.8196; ECGFounder 0.8228; HuBERT-ECG 0.8182. AUPRC ~0.26–0.28 for aortic and mitral endpoints	Single site; CIs not reported in the text reviewed; no calibration; small absolute differences; preprint
Lunelli et al., 2026 (BenchECG) [25]; I, with disclosed interest (authors developed xECG, a compared model; no competing interests declared); peer-reviewed (accelerated article preview)	5 FMs with released weights: ST-MEM, ECG-JEPA, ECG-CPC, ECGFounder, xECG (authors'). Supervised xLSTM without pretraining (ablation)	As released: HEEDB ~10M ECGs (ECGFounder, ECG-CPC); CODE-15%, Chapman and Ningbo ~200k (ST-MEM, ECG-JEPA); CODE, Chapman, Ningbo and IN-CART ~8M (xECG). Parameters 2.9M (ECG-CPC) to 85.4M (ECG-JEPA)	10 tasks on 8 public datasets: multilabel diagnosis (PTB-XL, CPSC2018), beat-level arrhythmia and R-peaks (MIT-BIH), sleep apnoea segmentation, PPG AF, exercise R-peaks, age, blood tests, survival; F1, F3, F5	Public datasets not listed among the compared models' pretraining data; CODE-15% (pretraining source for three models) used only for fine-tuning; age and survival fine-tuned on CODE-15% and tested unchanged on other populations (Supplementary Note 4)	Fine-tuning and linear probing; five runs per task with different seeds; task-specific hyperparameters reported in the article's supplement	BenchECG score (fine-tuning): xECG 0.838; ECG-CPC 0.797; ST-MEM 0.793; ECG-JEPA 0.783; supervised xLSTM 0.769; ECGFounder 0.588. Linear probing: xECG 0.771, others 0.550–0.655. PTB-XL AUROC 0.923–0.930 after fine-tuning. Within xECG, a smaller pretraining corpus lowered the score (0.820 to 0.802, signal-level views)	Run by developers of a compared model; pretraining corpora differ, so architecture and pretraining effects are confounded; threshold-free metrics; no calibration; equal task weighting in the score

Study [ID]; evaluator; status	Models compared	Pretraining data	Task(s); family	Evaluation data; separation	Adaptation; tuning	Key results (95% CI where reported)	Comparability limitations
B. Controlled developer comparisons and method studies							
Al-Masud et al., 2026 [26]; D (same group as [5]); preprint	Five self-supervised objectives (incl. CPC, JEPA, data2vec) on S4; Transformer and Net1D backbones	18k to ~11M ECGs	Multiple downstream benchmarks; mainly F1	Public benchmark datasets	Common backbone; continued pretraining on target-domain data tested	Pretraining loss followed a power law (exponent 0.189 CPC, 0.062 JEPA); downstream gains inconsistent; CPC best, JEPA close, data2vec lagged; S4 > Transformer and Net1D; continued pretraining helped	Developer protocol; loss scaling not equivalent to clinical gain; preprint
Nolin-Lapalme et al., 2026 (DeepECG) [6]; D; peer-reviewed	DeepECG-SSL vs DeepECG-SL (same group)	SSL: ~1.9M ECGs (institutional + CODE-15% + MIMIC-IV-ECG training split). SL: ~1.02M institutional ECGs	Interpretation (F1); LVEF ≤40% (F2); long QT detection and genotype; 5-year incident AF (F3)	Internal (patient-level 70/10/20); external public (MIMIC-IV-test patient-disjoint from its pretraining partition; CLSA, PTB, UK Biobank); seven external private health systems	Fine-tuning. SL: Bayesian sweep of 670 runs; SSL tuning not reported; per-label thresholds from the development validation split	Interpretation micro-AUROC SSL vs SL: 0.990 vs 0.992 (internal), 0.981 vs 0.980 (public), 0.983 vs 0.983 (private). LVEF ≤40%: 0.926 vs 0.917 (abstract and results) or 0.928 vs 0.900 (structured summary), n = 25,252. Random initialization (DeepECG article, Table S37): LVEF <50% 0.909 (supervised architecture) vs 0.885 and 0.887 fine-tuned; LVEF ≤40% 0.864 vs 0.917 and 0.926. The main text gives LQTS pretrained vs random initialization 0.875 vs 0.602, but the DeepECG article's Table S37 labels both values as randomly initialized models for LQTS detection; not used. LQTS detection 0.767 vs 0.735 (P = 0.117, n = 934); genotype 0.931 vs 0.850 (P = 0.026). 5-y AF 0.742 vs 0.734. External infarction/ischaemia: SL 0.867 vs SSL 0.863	Pretraining data differ in size and diversity; unequal tuning; LVEF values differ between results and structured summary; genotype sample size reported as 127 or 334; P values not adjusted for multiplicity
Yang et al., 2026 (ECG-Fyler) [8]; D; preprint	ECG-Fyler; ResNet from scratch; ViT masked autoencoder (same paediatric data); ECGFounder; EchoMini (external)	Paediatric ECG archive	LVEF ≤40% (F2)	Internal paediatric; external adult cohort	Fine-tuning with 100 labelled samples to full data; released checkpoints on same splits; comparator tuning not described	Internal, 100 samples, AUROC: Fyler 0.80 (0.79–0.80); scratch 0.50 (0.49–0.52); ViTMAE 0.63 (0.62–0.64); ECGFounder 0.47 (0.46–0.49). Full data: Fyler ~ scratch. External adult: 0.83 (abstract) / 0.84 (results); 100 external samples 0.81 (0.80–0.82) vs ECGFounder 0.72 (0.70–0.73). <10% external data: 0.87 vs EchoMini 0.85 (LVEF ≤45%)	Below-chance comparator unexplained; adaptation, input-handling or fine-tuning instability are untested hypotheses; EF threshold for the 0.81 vs 0.72 comparison not specified; preprint

Study [ID]; evaluator; status	Models compared	Pretraining data	Task(s); family	Evaluation data; separation	Adaptation; tuning	Key results (95% CI where reported)	Comparability limitations
Bollmann et al., 2026 [27]; D; conference abstract	Transformer time-series FM vs ResNet-18	33k unlabelled ECGs (single centre)	LV systolic dysfunction (F2)	31,832 ECG-echo pairs, single centre; no external validation	Not described in abstract	AUROC 0.924 (0.918–0.928) vs 0.912 (0.904–0.916), P < 0.001; AUPRC 0.695 vs 0.646 (not significant)	Abstract only; limited methods; single centre
C. Developer evaluations of individual foundation models							
Li et al., 2025 (ECG-Founder) [13]; D; peer-reviewed	ECGFounder vs CTN and S12L-ECG (CODE-test); ECG-SimCLR and ResNet (fine-tuning)	HEEDB, 10,771,552 ECGs from 1,818,247 subjects (whole database), 150 labels; after exclusions, development 1,319,128 and test 146,570 patients (published article; split method not stated)	Interpretation (F1); LVEF (F2); NT-proBNP, CKD, CHD (F3-F4); age, sex (F5); PPG AF	CODE-test, PTB-XL (overlapping classes), MIMIC-IV-ECG tasks, PhysioNet 2017	Frozen and fine-tuned	CODE-test mean AUROC 0.981 vs 0.980 (S12L-ECG), 0.963 (CTN). PTB-XL average 0.924 (text; the Sept 2025 correction replaced the PTB-XL table and the average was not re-derived). Fine-tuned: NT-proBNP 0.850, CKD 0.746, LVEF 0.867, CHD 0.795, age 0.861, sex 0.898; +2–3 points vs ECG-SimCLR, +4–6 vs ResNet. Single lead (PhysioNet 2017): NSR 0.975, AF 0.957	Developer evaluation; parity with best supervised comparator on CODE-test; PTB-XL limited to overlapping classes
McKeen et al., 2025 (ECG-FM) [16]; D; peer-reviewed	ECG-FM vs ResNet with multi-head attention and SE-WRN baselines	~1.5M ECGs, public datasets incl. MIMIC-IV-ECG, PTB-XL, CPSC2018	Interpretation (F1); LVEF ≤40% (F2)	UHN external cohort	Linear probing and full fine-tuning across training-set sizes	AF AUROC 0.996; LVEF ≤40% 0.929. Linear probe best at smallest sizes, then plateaus; full FT > baselines at ~100k samples; FT and baselines converge at large sizes	No CIs in text; 5-s inputs; weaker for MI and signal-quality labels
Coppola et al., 2024 (HuBERT-ECG) [10]; D; preprint	HuBERT-ECG small/base/large vs scratch; literature values for ECGFounder and specialized models	~9.1M ECGs; includes non-test partitions of evaluation datasets	Interpretation (F1); mortality (F3); paediatric; single lead	Multiple public datasets; test partitions excluded from pretraining	Fine-tuning	+3–8 AUROC points vs scratch on small datasets. vs ECGFounder: Ningbo 0.944 vs 0.953; ZZU-pECG 0.865–0.877 vs 0.889; MITHSDB 0.878–0.917 vs 0.965; SaMi-Trop mortality 0.871 (small) vs 0.702. CPSC 0.945–0.949 vs 0.965–0.980 (specialized). CPSC-Extra: small 0.894 vs base 0.756, large 0.774	Comparators copied from literature with different label scopes; pretraining overlaps evaluation datasets; preprint
Na et al., 2024 (ST-MEM) [28]; D; peer-reviewed conference	ST-MEM vs other self-supervised methods (e.g., MTAE, MLAE)	188,480 ECGs	Interpretation (F1)	PTB-XL, CPSC2018	Linear probe and fine-tuning; reduced-lead tests	PTB-XL linear 0.838 vs 0.807; FT 0.933 vs 0.915. CPSC linear 0.938 vs 0.818; FT 0.980 vs 0.973. 1% labels: PTB-XL 0.815, CPSC 0.897. Leads (PTB-XL): 12-lead 0.933, 6-lead 0.903, lead II 0.856, lead I 0.804	Comparators are self-supervised only; no tuned supervised baseline

Study [ID]; evaluator; status	Models compared	Pretraining data	Task(s); family	Evaluation data; separation	Adaptation; tuning	Key results (95% CI where reported)	Comparability limitations
Kim, 2024 (ECG-JEPA) [9]; D; preprint	ECG-JEPA vs ST-MEM and other self-supervised methods	~174k public 12-lead ECGs	Interpretation (F1)	PTB-XL, CPSC2018	Linear probe and fine-tuning	PTB-XL linear 0.912 vs 0.896 (ST-MEM); FT 0.944 vs 0.929. CPSC linear 0.971 vs 0.964; FT 0.983 vs 0.977	Some baselines retrained, others copied; released ST-MEM checkpoint; pretraining compute not matched (100 vs 300–800 epochs); preprint
Liu et al., 2024 (MERL) [29]; Wang et al., 2025 (MELP) [30]; D; peer-reviewed conferences	ECG-language models vs self-supervised ECG-only models	MIMIC-IV-ECG ECG-report pairs	Zero-shot and linear-probe classification (F1)	6 public datasets	Zero-shot; linear probe at 1–100% labels	MERL zero-shot mean AUROC 75.2%, 3.2 points above SSL linear probe at 10% labels. MELP zero-shot 79.0 vs MERL 75.3; PTB-XL rhythm linear 1%: 88.83 vs ECG-FM 81.45	Zero-shot well below supervised performance on the same tasks; comparators self-supervised, no tuned supervised baseline
Ko et al., 2026 (ECG-CLIP) [31]; D; peer-reviewed (abstract read)	ECG-CLIP vs two supervised baselines, a non-ECG foundation model, ECGFounder, DeepECG-SL and DeepECG-SSL	>1.7 million ECGs with clinician-overread reports from 542,288 patients (Scripps, 2008–2019); masked reconstruction then ECG-text contrastive training	AMI, amyloidosis, HCM (F2); incident AF (F3); other tasks not extracted	MIMIC-IV (>800,000 ECGs), five-fold cross-validation	Few-label adaptation by the developers for all models; details not read	With 10 positive labels: AMI 0.910 (0.903–0.916) vs 0.884; amyloidosis 0.790 vs 0.777; HCM 0.772 vs 0.754; AF prediction 0.777 vs 0.765 (best comparator in each case). On average 90.8% less training data to match the best comparator in detection tasks	Abstract only; comparators adapted by the developers of a competing model; DeepECG-SSL pretraining included a MIMIC-IV training partition; provisional patent held by four authors. Earlier secondary-report values removed
Hu et al., 2025 [32]; D; peer-reviewed	Pretrained deep model vs comparator for rare diseases	Institutional 12-lead ECGs (supervised pretraining at scale)	Rare structural disease (F2)	Institutional	Transfer learning with few labels	AUROC: carcinoid heart disease 0.772 vs 0.492; constrictive pericarditis 0.883 vs 0.689; rheumatic mitral disease 0.826 vs 0.701	Developer evaluation; rare-disease case numbers small; comparator details in original study
Diamant et al., 2022 (PCLR) [33]; D; peer-reviewed	Patient-contrastive pretraining vs training from scratch	Institutional 12-lead ECGs	4 tasks (F1-F3, F5)	Institutional	Linear models on embeddings vs scratch, 6 training sizes	Average improvement 51% relative to scratch across tasks and sizes	Comparator is scratch, not tuned supervised model
Hsieh et al., 2026 [20]; D; peer-reviewed	Transformer ECG foundation model (ECGFM) fine-tuned under sex-skewed, race-skewed and balanced training; no comparator model in the abstract	983,200 ECGs (self-supervised)	LV systolic dysfunction (F2)	Internal; external EchoNext	Fine-tuning on 44,815 ECG-echo pairs; training restricted to one sex or race group in skewed scenarios	AUROC 0.85–0.90 in both sexes after single-sex training; ≥ 0.90 across racial groups after White-only or non-White training; external EchoNext 0.823–0.917 by sex and 0.822–0.917 by race. Test cohort 8,963 patients	No from-scratch or supervised comparator in the abstract, so any contribution of pretraining to demographic robustness is not isolated; calibration not reported; authors probably the model developers

Study [ID]; evaluator; status	Models compared	Pretraining data	Task(s); family	Evaluation data; separation	Adaptation; tuning	Key results (95% CI where reported)	Comparability limitations
Xiao et al., 2026 [34]; D; peer-reviewed	Fine-tuned ECG FM	As released	Coronary stenosis on CT angiography (F2)	Internal and external cohorts	Fine-tuning	AUROC by vessel: internal 0.794, 0.818, 0.744, 0.755; external 0.749, 0.971, 0.667, 0.727	Comparator not identified in our extraction; wide external spread

Evaluator: I, independent of the developers of the evaluated models (competing interests disclosed, not disqualifying); D, developers of the model(s) or method. F1-F5, clinical task families (Section 2). AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; ARI, adjusted Rand index; FT, fine-tuning; LoRA, low-rank adaptation; LVEF, left ventricular ejection fraction; SSL, self-supervised learning; SL, supervised learning. Values are as reported by the cited study; comparator values in developer studies may come from other publications. Results from different rows are not directly comparable.

Table S2. Studies of generalization, robustness and reliability reviewed in Section 4, with evaluator and data source reported separately.

Study [ID]; status	Evaluator	Source of evaluation data	Model; adaptation; threshold	Endpoint; shift tested; validation cohort	Key results (95% CI where reported)	Limitations
A. Foundation-model-based or pretrained models						
Nolin-Lapalme et al., 2026 (DeepECG) [6, 7] (the second is a preprint); peer-reviewed (published full text read, 21 Oct 2026)	Developers with collaborating centres (private-site investigators are co-authors)	External to the development hospital: public CLSA, UK Biobank, PTB-XL and a 30% MIMIC-IV partition separated by patient from the pretraining partition (same source); private sets from seven US and Canadian centres, not pretraining sources	Self-supervised model pretrained on development-hospital, Code-15 and MIMIC-IV training ECGs, then fine-tuned on labelled development-hospital ECGs; supervised comparator from the same group. Per-label thresholds optimized on the development validation split and applied unchanged to all cohorts	77-condition ECG interpretation; micro-averaged AUROC (published abstract); internal vs external public vs external private; age and sex TPR/FPR differences	Micro-AUROC SSL vs SL: internal 0.990 vs 0.992; external public 0.981 vs 0.980; external private 0.983 vs 0.983 (published abstract gives CIs, zero width at 3 decimals). TPR differences < 0.1 and FPR differences < 0.02, micro-averaged across labels (published article; < 0.01 in the preprint). Weighted Brier score lower with SSL on PTB and MIMIC-IV (0.013 each) and UK Biobank (0.006), equal on CLSA, higher on the internal test set (0.078 vs 0.072); group-level rates only in an image figure not available to us	Dataset counts and some values differ between parts of the article; micro-average dominated by common labels; calibration only as a summary score; no device-stratified or temporal results found
Tang et al., 2026 (Pocket-K) [15]; preprint (full text read)	Developers (group includes the ECGFounder senior author)	Development: Peking University People's Hospital. Temporal: same hospital after July 2021, development patients excluded. External: second hospital in Tianjin, recorded 2024. Pretraining overlap not stated in our extract	ECGFounder initialization, fine-tuned on lead I (binary cross-entropy, Adam, learning rate 1e-4, up to 30 epochs, best validation AUROC); patient-level 8:1:1 split. Prespecified frozen threshold, value not reported	Hyperkalaemia, K > 5.5 mmol/L (secondary K ≥ 6.0 in Methods, '> 6.0' in abstract); ECG paired with nearest non-haemolysed potassium within 1 h. Development 10,409 patients, 26,145 pairs; temporal 5,054 patients, 10,922 pairs; external 18,976 patients, 25,223 pairs	AUROC K > 5.5: internal 0.9364 (0.9135–0.9575); temporal 0.8582 (0.8441–0.8723); external 0.8076 (0.7883–0.8258). K ≥ 6.0: temporal 0.9399 (0.9160–0.9590); external 0.8613 (0.8243–0.8940). Sensitivity / specificity internal 83.33% (78.24–87.43) / 96.60% (96.06–97.07); temporal sensitivity 85.05% (81.84–87.77); external NPV > 99.3%	Preprint; prevalence, calibration, threshold value and from-scratch comparator not reported; temporal and external steps differ in period, equipment (GE/Philips vs Nalong/Nihon Kohden; 500 vs 500 or 1000 Hz) and ECG-to-potassium interval; both cohorts in China; feasibility only for handheld device

Study [ID]; status	Evaluator	Source of evaluation data	Model; adaptation; threshold	Endpoint; shift tested; validation cohort	Key results (95% CI where reported)	Limitations
Lee S. et al., 2026 [14]; peer-reviewed	Developers	Second Korean hospital (Hospital B), external to the development hospital; relation to pretraining data not stated in our extract	Pretrained on machine-labelled ECGs, then fine-tuned on ECG-echo pairs; not described as an FM. Threshold: Youden index on development validation data	LVEF ≤ 40%; external hospital; follow-up pairs. Hospital A: 59,433 patients with echo, split by patient 8:1:1; Hospital B: 52,787	12-lead AUROC 0.953 (0.941–0.965) internal; 0.947 (0.943–0.951) external. Follow-up pairs before adjustment: 0.945 vs 0.910 (0.905–0.916); after patient-level adjustment: 0.956 vs 0.940 (0.936–0.945). Specificity at Youden threshold, internal vs external: 0.852 vs 0.763 before; 0.912 vs 0.859 after	No temporal validation (random patient split); no calibration slope, intercept or Brier score; adjustment uses prior echo results and is not a robustness measure; two Korean hospitals
Hsieh et al., 2026 [20]; peer-reviewed	Developers (probable)	Test cohort from the same health system; external EchoNext cohort (public). Overlap with pretraining not stated in abstract	Transformer FM pretrained on 983,200 ECGs; fine-tuned on 44,815 ECG-echo pairs under sex-skewed, race-skewed and balanced training. Thresholds not described in abstract	LV systolic dysfunction; demographic imbalance in training; test cohort 8,963 (52% male); external EchoNext	Single-sex training: AUROC 0.85–0.90 in both sexes. White-only or non-White training: ≥ 0.90 in all racial groups. External EchoNext: 0.823–0.917 by sex, 0.822–0.917 by race	No from-scratch comparator in abstract; calibration not reported; one task; CIs not in abstract
Xiao et al., 2026 [34]; peer-reviewed	Developers	External cohort (institution not extracted)	Fine-tuned ECG FM; adaptation and threshold not extracted	Coronary stenosis on CT angiography, four vessels; internal and external cohorts (sizes not extracted)	AUROC by vessel, internal: 0.794, 0.818, 0.744, 0.755; external: 0.749, 0.971, 0.667, 0.727	Sample sizes, CIs and comparator not in our extract; wide external spread, one vessel higher externally
Yang et al., 2026 (ECG-Fyler) [8]; preprint	Developers	Adult cohort at CUIMC, external to paediatric development; test patients excluded from pretraining	Paediatric-pretrained FM; evaluated without target fine-tuning; thresholds not extracted	LVEF ≤ 50%, ≤ 40%, ≤ 30% (adult ECG-echo pairs: 82,543, test 5,442); paediatric lesion subgroups	AUROC 0.81, 0.84, 0.87 for LVEF ≤ 50%, ≤ 40%, ≤ 30%. Tetralogy of Fallot 0.50 (0.44–0.57) to 0.81 (0.74–0.86); paced rhythm 0.47 (0.42–0.51) to 0.61 (0.56–0.65) with pretraining	No sex, age, race, device or site subgroups; no calibration, temporal or robustness analyses; device and sampling rate of external cohort not reported; patient-level splitting not explicit
McKeen et al., 2025 (ECG-FM) [16]; peer-reviewed (main text read; supplement not seen)	Developers (all authors from the developing institutions; no competing interests declared)	UHN-ECG: developers' own institutions, excluded from pretraining. MIMIC-IV-ECG labels: public, also a pretraining source	Pretrained on public ECGs (1,405,625 samples). Arms: full fine-tuning (learning rate 1e-6), linear probe, random initialization, two supervised baselines. Thresholds set on validation data for a target recall; best validation AUPRC checkpoint	Multi-label interpretation and reduced LVEF (≤ 50, 40, 35, 30%; echo within 7 days). UHN-ECG 622,000 ECGs, 211,000 patients, 2010–2018, split 80/10/10 by patient and time (simulated prospective); training-data fractions 100, 50, 10, 1%	Abstract: AUROC 0.996 for atrial fibrillation and 0.929 for LVEF ≤ 40%, dataset not named in the abstract. Per-task results and scaling in supplementary tables or a figure, not seen. No CIs	Not an external validation of a locked model; adapted on the same institution's training split; MIMIC-IV-ECG task shares a source with pretraining; no calibration, subgroup, device, lead-dropout or test-retest analysis in main text; models not shareable

Study [ID]; status	Evaluator	Source of evaluation data	Model; adaptation; threshold	Endpoint; shift tested; validation cohort	Key results (95% CI where reported)	Limitations
Monachino et al., 2026 [17, 18] (evaluation publication and backbone reference, respectively); peer-reviewed (abstract and supplementary information read; main text not read)	Developers (same group as the Self-DANA backbone)	Unseen in-domain and out-of-domain public datasets; per the supplementary information, small public sets	Self-DANA backbone. Compared full fine-tuning with partial fine-tuning, BitFit, LayerNorm tuning and linear probing. Tuning budget and seeds not found	Adaptation strategy and out-of-domain generalization across tasks and dataset characteristics (tasks and sizes not extracted)	Abstract: a well-chosen selective strategy outperformed full fine-tuning in resource efficiency and generalization, especially on out-of-domain data. No numbers extracted	One backbone; developers; best strategy depended on task (supplementary information); main text not read; seeds and tuning budget not found
Ko et al., 2026 (ECG-CLIP) [31]; peer-reviewed (abstract read)	Developers (Scripps Research authors)	MIMIC-IV (>800,000 ECGs) as external dataset; comparator DeepECG-SSL was pretrained on a MIMIC-IV training partition [6]	Contrastive ECG-text FM; five-fold cross-validation on MIMIC-IV, so adapted rather than locked; comparators include ECGFounder and both DeepECG models	Disease detection with few positive labels; incident AF	AMI AUROC 0.910 (0.903–0.916) vs 0.884 with 10 positive labels (other values in Supplementary Table S1)	Abstract only; no locked external validation; evaluators developed the model
Li et al., 2025 (ECGFounder) [13]; peer-reviewed (author manuscript of the published article read)	Developers	HEEDB development; CODE-test, PTB-XL, PhysioNet 2017 outside HEEDB. Overlap with pretraining not described in the text read	Supervised-at-scale FM; lead-augmented single-lead variant; missing leads set to zero	Cross-region test; single lead. HEEDB development 7.5M ECGs from 1.32M patients; test 0.83M ECGs from 0.15M patients (published; the arXiv version gave 1.73M and 0.27M patients; 1.82M is the whole database)	CODE-test AUROC 0.981; PTB-XL average 0.924 (overlapping classes; stated in text; the Sept 2025 correction replaced the PTB-XL table, average not re-derived); PhysioNet 2017 single lead: sinus rhythm 0.975, AF 0.957	No subgroup, calibration, noise, sampling-rate or manufacturer analysis in text read; split method of HEEDB not stated; patient counts differ between the arXiv and published versions
Na et al., 2024 (ST-MEM) [28]; peer-reviewed conference	Developers	Public (PTB-XL); relation to pretraining not extracted	Masked FM; lead-subset evaluation	Lead reduction; PTB-XL superclass	AUROC 0.933 (12 leads), 0.903 (6 limb leads), 0.856 (lead II), 0.804 (lead I)	Single dataset; no noise or device analysis
Coppola et al., 2024 (HuBERT-ECG) [10]; preprint	Developers	Public datasets; pretraining includes non-test partitions of evaluation datasets	Masked-prediction FM; comparator values partly from other papers	Single-lead and paediatric datasets	Single-lead (MITHSDB) AUROC 0.878–0.917 vs 0.965 for ECGFounder	Comparators not matched; preprint
Nguyen et al., 2025 (TolerantECG) [22]; preprint	Developers	PTB-XL; MIT-BIH Arrhythmia (public)	FM for noisy or incomplete ECGs; adaptation not extracted	Noise; missing leads; lead subsets	Reported as first or second across signal conditions on PTB-XL and first on MIT-BIH; no values in abstract	Baselines not named in abstract; noise types and severity not specified; preprint; claim cannot be evaluated

Study [ID]; status	Evaluator	Source of evaluation data	Model; adaptation; threshold	Endpoint; shift tested; validation cohort	Key results (95% CI where reported)	Limitations
B. Independent or partly independent benchmarks (externally sourced public data; in-distribution or adapt-and-test, except cross-population transport in BenchECG)						
Filice et al., 2026 [11]; preprint	Independent	Public: PTB-XL, CODE-15%, Georgia, Chapman-Shaoxing/Ningbo; overlap with pretraining not audited	Six frozen encoders with classical classifiers; 15-fold cross-validation	Dataset vs label structure of embeddings; feature overlap across datasets	Dataset ARI (conduction disturbance / AF): ECG-JEPA 0.98/1.00; HuBERT-ECG 0.42–0.47/0.66–0.71; ECG-FM 0.31/0.20; ECGFounder 0.03/0.12. Label ARI: ECG-JEPA 0.07/0.00; ECG-Founder 0.24/0.70	No calibration, noise, subgroup or train-on-one-test-on-another analysis; no supervised baseline; geometry does not establish dependence
Al-Masud et al., 2026 [5]; peer-reviewed conference	Partly independent	12 public datasets; several also in pretraining corpora; no overlap audit	8 FMs and 2 supervised models; adaptation as in Supplementary Table S1	In-distribution performance across seven categories	See Supplementary Table S1	No external transport; overlapping datasets; no overlap audit
Taleshinosrati et al., 2026 (FOUND-AF) [23]; preprint	Independent	Four public single-lead AF datasets (AFDB, CinC2017, CPSC2021, LTAFDB); sampling rates 128–300 Hz; devices not described	Frozen encoders with XGBoost	AF detection within each dataset; five-fold cross-validation grouped by recording (patient-level grouping not described); fixed threshold 0.5	See Supplementary Table S1	Within-dataset only; no supervised baseline or calibration; preprint
Do et al., 2026 [24]; preprint	Independent	EchoNext (public), single site	FMs and ResNet; adaptation strategies compared	Structural heart disease endpoints; precision-recall performance	Macro AUROC 0.82–0.85; AUPRC about 0.26–0.28 for aortic and mitral endpoints	Single site; no external or calibration results
Lunelli et al., 2026 (BenchECG) [25]; peer-reviewed (accelerated article preview)	Independent of four of the five compared models; authors developed xECG	Public. Age: fine-tuned on CODE-15% (Brazil), tested on MIMIC-IV-ECG (USA), PTB-XL (Germany), CPSC2018 (China). Mortality: CODE-15% (primary care) to MIMIC-IV-ECG (intensive care)	Five FMs and a supervised xLSTM; fine-tuned or linear-probed on the training dataset and applied unchanged to the test datasets; no threshold	Age (MAE, R^2) and all-cause mortality (C-index); population, care-setting and prevalence shift (5-year mortality 4.7% vs 36%)	Fine-tuned, test sets: age mean MAE 7.560–8.854 years (supervised xLSTM 8.896); C-index 0.677–0.710 (supervised xLSTM 0.626); C-index drop from validation 0.082–0.133 (supervised xLSTM 0.024)	Threshold-free tasks without a declared clinical use; no calibration; evaluators developed a compared model; not locked external validation of a clinical classifier
C. Task-specific AI-ECG models (context, not foundation-model evidence)						
Carter et al., 2026 [35]; peer-reviewed	Developer-affiliated (Mayo Clinic and other-site investigators, Anumana employees)	Four geographically diverse US sites; patients independent of development (per abstract)	Existing commercial algorithm; threshold not extracted	LVEF; 13,960 evaluable of 16,000 patients; mean age 66; 52% male; prevalence 7.9%	AUROC 0.92 (0.91–0.93); sensitivity 84.5% (82.2–86.6); specificity 83.6% (82.9–84.2); PPV 30.5% (28.8–32.1); NPV 98.4% (98.2–98.7)	Developer-affiliated authors; ECG-echo pairs within 30 days; retrospective

Study [ID]; status	Evaluator	Source of evaluation data	Model; adaptation; threshold	Endpoint; shift tested; validation cohort	Key results (95% CI where reported)	Limitations
Lee H. et al., 2026 [36]; peer-reviewed	Developers with Mayo Clinic investigators	Mayo Clinic Platform US data, external to the Korean developers	ECG Buddy (ten biomarkers); prespecified non-inferiority benchmarks	Emergency and cardiac-function cohorts; paired subset N = 1,368	AUC 0.883–0.949, all meeting prespecified non-inferiority criteria; consistent across sex and age strata; AMI 0.920 vs troponin T 0.878	Calibration deferred to prospective work; retrospective; developer co-authors
Sau et al., 2024 (AIRE) [37]; peer-reviewed	Developers	External cohorts in USA, Brazil, UK	Survival model trained at one centre (BIDMC) and applied externally	All-cause mortality; development 1,163,401 ECGs, 189,539 patients	C-index 0.775 (0.773–0.776) internal; 0.638–0.773 external (range across cohorts)	Developer-led; outcome models depend on follow-up and case mix
Noseworthy et al., 2020 [21]; peer-reviewed	Developers (Mayo Clinic)	Separate test cohort from the same health system	CNN trained on derivation cohort; threshold not extracted	Low LVEF by race and ethnicity; 97,829 patients; derivation 44,959 (96.2% non-Hispanic White); test 52,870	AUC 0.931–0.961 across five groups; White-only training ≥ 0.930; race classifier AUC 0.84 (Black), 0.76 (non-Hispanic White)	One task, one health system; subgroup sensitivity and PPV not given in abstract
Sangha et al., 2022 [38]; peer-reviewed	Developers	External test in Germany	Image-based multilabel CNN	Rhythm and conduction; image vs signal; 2,228,236 ECGs from Brazil, external 21,785 ECGs	AUROC 0.99 held-out; 0.97 external; image model beat signal model; 120 of 120 reviewed high-confidence false positives reported as label errors	Rhythm and conduction labels only; abstract read via reconstruction
Dhingra et al., 2025 [19]; preprint	Developers (declared conflicts of interest)	Five hospitals and outpatient network of one health system	Six CNNs and an ensemble (PRESENT-SHD)	Test-retest at 1–30 days; image formats; demographic effects on instability; 1,118,263 ECG pairs	CCC ensemble 0.90 (0.89–0.92 external), CNNs 0.80–0.88; kappa 0.75–0.82 vs 0.57–0.72. Instability OR: age < 65 2.52 (2.38–2.67); both inpatient 1.91 (1.78–2.05). Image formats Spearman 0.942–0.998	Interval ≥ 1 day; 'external' sites within one system; no ICC or flip rate; conclusions inconsistent; preprint
Scagnetto, 2026 [12]; preprint	Single author	MIMIC-IV-ECG and HEEDB (public)	ECG biometric model; not an FM	Identity signal across institutions and time; internal 164,440 ECGs, 53,079 patients	Rank@1 0.9506 (internal), 0.8291 (MIMIC), 0.6884 (HEEDB); 1 to 5 years: 0.7853 to 0.6433 (MIMIC), 0.6864 to 0.5560 (HEEDB)	Not an FM; split design not confirmed; full text not read

Evaluator: developers, the group that built the model (collaborators are co-authors from partner institutions); independent, no stake in the model; partly independent, mixed. Source of evaluation data is reported separately: external means outside the developers' institutions, and pretraining overlap is noted where stated. AUROC, area under the receiver operating characteristic curve; ARI, adjusted Rand index; CCC, concordance correlation coefficient; OR, odds ratio; PPV and NPV, positive and negative predictive value; TPR and FPR, true- and false-positive rate. Rows in section C concern task-specific models and are context only. Values are as reported by the cited studies, are not comparable across rows, and differences between internal and external values do not identify their cause.

Table S3. Clinical evidence for AI-enabled ECG by study design: randomized trials, a non-randomized interventional study, post hoc analyses, economic modelling and systematic reviews (conventional, task-specific algorithms; none is a foundation-model study).

Study [ID]; design	Population; setting; size	Intervention; comparator; threshold	Primary endpoint and result (95% CI)	Secondary, process and safety results	Limitations; interests; source read
A. Randomized trials					
EAGLE, Yao et al., 2021 [39]; pragmatic cluster-randomized (NCT04000087)	22,641 adults without prior heart failure, ECG in routine care (11,573 intervention, 11,068 control); 120 primary care teams, 358 clinicians; Mayo Clinic	Access to AI-ECG result; EHR recommendation to consider echocardiography plus email alert (per [40]); positivity threshold set for 9.0% prevalence of low EF, about 6% of ECGs positive; usual care control	New diagnosis of EF ≤ 50% within 90 days: 2.1% vs 1.6%, OR 1.32 (1.08–1.61) in the Results text, (1.01–1.61) in the abstract, P = 0.007 (see note). AI-positive patients: 19.5% vs 14.5%, OR 1.43 (1.08–1.91), P = 0.01	Echocardiography overall 19.2% vs 18.2% (P = 0.17); AI-positive 49.6% vs 38.1% (P < 0.001). Among tested AI-positive patients, EF ≤ 40% in 22.1% vs 22.7% (per [41]). Clinician response reported separately [40]	Algorithm developed at Mayo Clinic; no multiplicity adjustment for echo analyses; one health system, about 94% White (per [40]); patents and licensing interests; journal page (abstract, figure legends) read
Tsai et al., 2025 [42]; pragmatic randomized, individual randomization by medical record number (NCT05117970)	13,631 inpatients under non-cardiologist care (6,840 vs 6,791); one academic centre, Taiwan	AI-generated ECG results as decision support vs standard care; threshold for 'high risk' not extracted	New low EF (≤ 50%) within 30 days: 1.5% vs 1.1%, HR 1.50 (1.11–2.03); the abstract gives P = 0.023, which the figure legend gives as the P for interaction. High-risk patients: 13.0% vs 8.9%, HR 1.55 (1.08–2.21)	Echocardiography use 17.1% vs 17.3%, HR 1.00 (0.92–1.09). Echocardiograms finding low EF among high-risk patients 34.2% (79/231) vs 20.2% (50/248). 30-day mortality 4.3% vs 4.5%, HR 1.00 (0.84–1.17). Cardiology consults in high-risk patients 29.3% vs 23.5% (P = 0.027, post hoc)	Single centre; 30-day window; no treatment data after diagnosis; algorithm developed at the same institution; authors declare no competing interests; full text read (PMC)
Lin et al., 2024 [43]; pragmatic randomized (NCT05118035); randomization unit not stated in text read	15,965 hospitalized patients; 39 physicians; emergency and inpatient departments, Taiwan	AI-ECG report plus warning to physicians for high predicted mortality risk (95th percentile of score); non-alert control	All-cause mortality at 90 days: 3.6% vs 4.3%, HR 0.83 (0.70–0.99). High-risk ECG subgroup (prespecified analysis): HR 0.69 (0.53–0.90); P for interaction 0.026	Cardiac death in high-risk subgroup: 0.2% vs 2.4%, HR 0.07 (0.01–0.56). More intensive care among intervention patients with high-risk ECGs (numbers not in text read). Response to alerts not reported	Mechanism not established; subgroup defined by the algorithm; extended analyses not adjusted for multiplicity; algorithm licensed to a company, 13 authors may benefit financially; full text read
ARISE, Lin et al., 2024 [44]; described as two-centre open-label cluster-randomized (NCT05118009); allocation by date of first ECG (odd or even days), with assignment revealed daily to 20 on-duty cardiologists	43,234 eligible patients; emergency department and wards, academic and community hospital sharing one catheterization laboratory, Taiwan, May 2022 to April 2023; STEMI by angiography in 77 vs 68	AI-ECG-assisted STEMI detection vs standard care; threshold not extracted	Door-to-balloon time, ED, median: 82.0 min (IQR 62.5–89.5) vs 96.0 (78.0–137.0), P = 0.002	ECG-to-balloon (ED and inpatient) 78.0 vs 83.6 min, P = 0.011. PPV 89.5% (85.3–93.6), NPV 99.9% (99.9–100.0). New HFrEF 340 vs 304 (OR 1.12, P = 0.151); cardiac death 85 vs 116 (OR 0.73, P = 0.029); all-cause death 1,153 vs 1,127 (OR 1.02, P = 0.568)	Process endpoint; 145 confirmed cases; no CI for time difference; secondary outcomes uncorrected; two sites, one catheterization laboratory; date-based allocation; full text read; disclosure forms not read

Study [ID]; design	Population; setting; size	Intervention; comparator; threshold	Primary endpoint and result (95% CI)	Secondary, process and safety results	Limitations; interests; source read
Lin et al., 2026 (potassium alert) [45]; pragmatic open-label physician-level randomized (NCT05118022)	14,989 emergency patients (7,506 vs 7,483); 70 emergency physicians (35 vs 35); academic and community hospital, Taiwan; six months in 2022	Real-time pop-up alert for risk of K ≥ 6.0 or ≤ 3.0 mmol/L from AI-ECG vs no alert	Treatment within 3 h: hyperkalaemia 8.0% vs 7.7%, HR 1.05 (0.94–1.17), P = 0.420; hypokalaemia 2.1% vs 2.4%, HR 0.91 (0.74–1.13), P = 0.392	Among patients identified by AI-ECG as hyperkalaemic: treatment 69.1% vs 41.6%, HR 2.23 (1.44–3.46), P < 0.001	Process endpoint, no clinical outcome; subgroup defined by the algorithm; two sites; algorithm licensed by the developing university to a company; full text read
SPEC-AI Nigeria, Adedinsewo et al., 2024 [46]; pragmatic open-label randomized (NCT05438576)	1,232 pregnant and postpartum women randomized (616 per arm), 1,195 at baseline (587 vs 608); six hospitals, Nigeria	AI-guided screening with an AI digital stethoscope, or with a 12-lead AI-ECG (separate arm), vs usual care	LVSD detected: stethoscope 24/587 (4.1%) vs 12/608 (2.0%), OR 2.12 (1.05–4.27), P = 0.032; AI-ECG 20/587 (3.4%) vs 2.0%, OR 1.75 (0.85–3.62), P = 0.125	Similar direction in a prespecified subgroup analysis. No serious adverse events related to participation	Detection yield only; 12-lead result imprecise; the Feb 2025 correction adds an omitted acknowledgement and does not change results; full text read
TRICORDER, Kelshiker et al., 2026 [47]; pragmatic cluster-randomized implementation trial	205 UK primary care practices (96 vs 109); 701,933 vs 851,242 registered patients	AI stethoscope (15-s single-lead ECG plus phonocardiogram; algorithms for EF ≤ 40%, AF, valvular disease) with training vs routine care. Adjacent evidence: not a 12-lead ECG	Newly coded heart failure per 1,000 patient-years: IRR 0.94 (0.86–1.02)	12,725 examinations by 972 users. Use of the stethoscope was associated with higher detection of heart failure, AF and valvular disease after 12 months (not randomized). Clinician-reported barriers collected	Coded diagnoses from routine data; usage comparison confounded by who used the device; abstract-level source
DULCE, Simonetto et al., 2026 [48]; pragmatic cluster-randomized (NCT05782283)	15,596 adults with routine 12-lead ECGs (8,034 vs 7,562); 98 primary care teams (123 vs 122 clinicians); Mayo Clinic	ECG machine-learning result available to clinicians vs usual care; target undiagnosed advanced chronic liver disease (non-cardiac)	New advanced chronic liver disease within 180 days: 1.0% vs 0.5%, OR 2.09 (1.22–3.55), P = 0.007	ECG-ML-positive patients: 4.4% vs 1.1%, OR 4.37 (1.94–9.88). Any fibrosis 1.7% vs 0.5%, OR 3.17 (1.86–5.40). Yield below epidemiological estimates, attributed by the authors to variable adherence	Full text read; the Feb 2026 author correction amends an author's name only; algorithm licensed to a company and eight authors may benefit; 96% white population; harms not reported
B. Prospective non-randomized interventional study					
BEAGLE, Noseworthy et al., 2022 [49]; prospective non-randomized interventional	1,003 participants (mean age 74) with stroke risk factors and no known AF, from 40 US states (NCT04208971)	AI-ECG classified participants as high or low AF risk; all wore a continuous ambulatory monitor for up to 30 days (mean 22.3 days)	AF detected: 48/633 (7.6%) high risk vs 6/370 (1.6%) low risk, OR 4.98 (2.11–11.75), P = 0.0002	Versus propensity-matched eligible but unenrolled usual-care controls: high risk 10.6% vs 3.6% (P < 0.0001); low risk 2.4% vs 0.9% (P = 0.12); median follow-up 9.9 months	Risk-group comparison shows enrichment, not an intervention effect; usual-care comparison observational; Mayo-developed algorithm; abstract read (Mayo Clinic repository)

Study [ID]; design	Population; setting; size	Intervention; comparator; threshold	Primary endpoint and result (95% CI)	Secondary, process and safety results	Limitations; interests; source read
C. Post hoc or secondary analyses of trials					
Clinician adoption in EAGLE, Rushlow et al., 2022 [40]; secondary analysis of the intervention arm	165 clinicians (60 teams); 11,573 patients; Mayo Clinic primary care	Adoption = echocardiograms ordered per AI-positive ECG; high adopters = top quartile (n = 41, ≥ 64.3%)	Echocardiogram ordered for AI-positive ECGs: 78.7% (100/127) high adopters vs 43.0% (243/565) others	Low EF found per AI-positive ECG: 33.9% vs 16.3% (43/127; 92/565); among echocardiograms 43.0% vs 37.9%. Adjusted OR for detection 1.62 (1.21–2.17). NP/PA high adopters 37.7% vs physicians 18.8%; no clinician trait significant after adjustment	Reasons for not ordering not captured; single system; minor internal inconsistencies (yield 16.3% in text, 16.9% in abstract); counts agree with the trial report; Mayo patent interest; journal page text read
DULCE clinician engagement, Calleri et al., 2026 [50]; post hoc analysis of the intervention arm	110 clinicians who received at least one alert; 1,385 patients; Mayo Clinic primary care	Engagement = acting on ECG-ML recommendations; high engagement = top quartile	Advanced CLD detection, high vs low engagement: OR 2.12 (1.36–3.30)	Overall engagement 29.8%. Diagnostic yield for advanced CLD 10.6% vs 2.9%, OR 2.99 (1.73–5.16); any CLD 22.3% vs 5.0%	Post hoc; engagement not randomized; abstract (via OpenAlex) read; article number not retrieved
VITAL-AF risk-guided analysis, Vedage et al., 2026 [51]; analysis of a cluster-randomized screening trial (NCT03515057)	30,630 aged ≥ 65 without prevalent AF; 16,937 with a pretrial ECG and clinical data; 16 primary care practices (MGH)	Single-lead ECG screening vs usual care; risk from CHARGE-AF, a 12-lead ECG AI model (ECG-AI) and combined (CH-AI) models developed outside the trial	Two-year AUROC: CHARGE-AF 0.711 (0.671–0.749); ECG-AI 0.784 (0.743–0.819); CH-AI 0.788 (0.754–0.824)	Top CH-AI decile: 10.07 vs 7.76 AF diagnoses per 100 person-years; difference 2.32 (0.01–4.63); number needed to screen 43 per year	Effect estimated within risk deciles, post hoc; interval reaches zero; detection is a surrogate for stroke prevention; abstract read
VITAL-AF single-lead prediction, [52]; prediction analysis within a randomized trial	15,694 without prevalent AF (2-year incident AF 3.1%); test set 4,221	Published single-lead ECG AI model with age and sex vs CHARGE-AF; no intervention tested	AUROC 0.695 vs 0.679 (P = 0.46)	Net reclassification vs age ≥ 65: 0.27	Discrimination only; abstract read
D. Economic modelling based on trial inputs					
Cost-effectiveness of EAGLE, Thao et al., 2024 [41]; model based on trial inputs	Model population 22,641 (median age 63); payer perspective, 70-year horizon; costs in 2020 USD	AI-ECG cost $100; echocardiography $1,801; decision tree feeding a lifetime Markov model	ICER $27,858 per QALY (incremental cost $386; incremental QALYs 0.0139)	Outpatient ICER $1,651; scenarios $13,007 (age 20) to $36,213 (age 80); cost-effective in 79.5% of simulations at $100,000 per QALY	Modelled, not observed outcomes; EF ≤ 40% only; single health system; authors with patents and licensing; PMC text read (DOI 10.1016/j.mcpdig.2024.10.001)
E. Systematic reviews of randomized trials (AI in cardiovascular care, all modalities)					
Systematic review and meta-analysis [53]	31 RCTs (1,685,717 patients); 11 in meta-analysis; searches to 28 April 2026	Any AI intervention in cardiovascular care vs routine care	22 of 31 (71%) reported a significant primary benefit, mostly intermediate process measures; 9 (29%) no effect	AI-enhanced ECG, heart failure detection: RR 1.22 (0.70–2.14), I2 90%. Image-based decision support, MACE: RR 0.74 (0.58–0.96)	7 of 31 at low risk of bias; high heterogeneity; ECG estimate very imprecise; does not isolate foundation models

Study [ID]; design	Population; setting; size	Intervention; comparator; threshold	Primary endpoint and result (95% CI)	Secondary, process and safety results	Limitations; interests; source read
Systematic review [54]	11 RCTs, 2021 to 2024, 81.2% multicentre (as stated)	Machine-learning models vs traditional care	5 of 11 (45.5%) improved clinical events; 6 (54.5%) improved diagnostic accuracy or early detection	3 of 11 (27.3%) improved resource use	Risk of bias not reported in the abstract; abstract-level source
F. Clinical evidence for ECG foundation models					
No trial, prospective clinical evaluation or implementation study identified [15]	A single-lead hyperkalaemia preprint reports a handheld prototype on two examples only	Not applicable	Not applicable	Prototype accuracy not evaluated	Statement limited to the literature reviewed; registry search in Supplementary Table S12

All studies concern conventional AI-ECG algorithms except where noted. CLD, chronic liver disease; ECG-ML, ECG machine-learning model. OR, odds ratio; HR, hazard ratio; IRR, incidence rate ratio; RR, risk ratio; EF, ejection fraction; LVSD, left ventricular systolic dysfunction; PPV and NPV, positive and negative predictive value; ICER, incremental cost-effectiveness ratio; QALY, quality-adjusted life-year; AF, atrial fibrillation; STEMI, ST-elevation myocardial infarction. Values are as reported by the cited sources and are not comparable across rows. Where only an abstract was read this is stated. Our arithmetic is labelled in the text. EAGLE: OR 1.32 (1.08–1.61), P = 0.007, in the Results text and Supplementary Table 6 (mixed-effect logistic regression, care team as random effect); 1.01–1.61 in the abstract and PubMed record; no correction published; discrepancy unresolved. Full texts of EAGLE, the inpatient, mortality-alert, potassium, SPEC-AI, DULCE and ARISE trials were read on 21 October 2026.

Table S4. Lessons from conventional AI-ECG trials and whether they transfer to ECG foundation models (our reasoning, not trial findings).

Lesson	Evidence from conventional AI-ECG studies	Does it transfer to foundation models?	What a foundation-model study would need to show
Endpoints form a hierarchy; process gains are not patient benefit	Most primary endpoints in 31 trials were intermediate [53]; low-EF diagnoses rose in three trials without reported treatment or outcome data [39,42,46]	Yes, as a design principle for any model	Pre-specified endpoint level; process and outcome results reported separately
Thresholds fix the alert burden and predictive values	About 6% of ECGs positive in EAGLE; top 5% flagged in the mortality trial [39,43]; 135 of 343 echocardiograms after AI-positive ECGs found low EF (our arithmetic) [40]	Yes; adaptation at each site can move the threshold (Section 4)	Threshold, alert rate, PPV and yield per confirmatory test at the target prevalence
Clinician response determines the effect	Echocardiogram ordered for 78.7% vs 43.0% of AI-positive ECGs [40]; yield 10.6% vs 2.9% by clinician engagement in DULCE [50]; null overall effect where use was low or alerts broad [45,47]	Yes; a better backbone does not act on this step	Response rate, reasons for not acting, and effect by response level
Added diagnoses may reflect more tests or better selection	EF ≤ 40% in 22.1% vs 22.7% of tested AI-positive patients, with more testing, in EAGLE [39,41]; unchanged echocardiography with higher yield among high-risk patients in the inpatient trial [42]	Yes	Whether gains come from selection or volume, relative to the incumbent pathway

Lesson	Evidence from conventional AI-ECG studies	Does it transfer to foundation models?	What a foundation-model study would need to show
Algorithm-defined subgroups overstate effect	AI-positive groups in EAGLE, the potassium trial and the mortality trial [39,43,45]	Yes	Subgroups defined by model output labelled exploratory; effect in the whole eligible population
Population and equity are under-reported	EAGLE 94% and DULCE 96% White [39,48]; no effects by race or ethnicity found in the sources read	Yes; foundation-model subgroup evidence is itself limited (Section 4)	Subgroup performance at the operating threshold and effects by group
A trial effect belongs to the exact algorithm, threshold and workflow	Each trial tested one package; results were heterogeneous across settings [53]	No: effect estimates do not transfer to a different model	Bridging conditions of principle P7 (Section 6, Table 4); a presumption at most, ranked below a trial of the successor (our proposal)
Multi-output platforms multiply alerts	TRICORDER combined three algorithms and did not raise heart failure detection overall [47]; conventional, not a foundation model	Unknown: hypothesis that a shared representation changes alert burden is untested	Per-output thresholds and false-positive burden for the full output set

FM, foundation model; PPV, positive predictive value. Right-hand columns are our reasoning and proposals; they are not findings of the cited studies.

Table S5. Evidence matrix: three ECG foundation-model backbones (Level 1) and five worked uses (Levels 2 and 3). Cell codes: state, evaluator, preprint and context marks. Consensus codes after independent double coding and adjudication (Supplementary Note 5); both raters' original codes are in Supplementary Data 2.

ID	Evidence requirement	ECGFounder (backbone)	ECG-FM (backbone)	DeepECG-SSL (backbone)	Hyperkalaemia, single lead (ECG-Founder, fine-tuned)	AF detection, single lead (ECG-Founder, frozen)	Structural heart disease (ECG-FM, adapted)	77-condition interpretation (DeepECG-SSL, fine-tuned)	Reduced LVEF (DeepECG-SSL, fine-tuned)
Level 1. Backbone									
B1	Pretraining data separated from evaluation data (disjointness or overlap audit stated)	U·D p	V·D	V·D+C					
B2	Reuse across two or more clinically distinct task families, with non-ECG reference standards	V·D+I	V·D+I	V·D					
B3	Label efficiency against a supervised comparator on the same labels	V·I	V·D	U·D					
B4	Backbone evaluated by investigators other than its developers	V·I	V·I	U·I					
B5	Embedding analysis: source dataset versus label structure	V·I p	V·I p	N					
B6	Interventional test of reliance on acquisition features	U·I p	U·I p	N					

ID	Evidence requirement	ECGFounder (backbone)	ECG-FM (backbone)	DeepECG-SSL (backbone)	Hyperkalaemia, single lead (ECG-Founder, fine-tuned)	AF detection, single lead (ECG-Founder, frozen)	Structural heart disease (ECG-FM, adapted)	77-condition interpretation (DeepECG-SSL, fine-tuned)	Reduced LVEF (DeepECG-SSL, fine-tuned)
B7	Prompt sensitivity of zero-shot use	X	X	X					
Level 2. Adapted model									
A1	Discrimination against a supervised comparator, same labels, target population				N	N	V·D p	U·D+C	U·D+C
A2	Evaluation patients stated to lie outside the pretraining corpus				U·D p	V·D p	U·D p	V·D+C	V·D+C
A3	Patient-level separation of adaptation and test partitions stated				V·D p	U·D p	U·D p	V·D	V·D
A4	Locked external validation (fixed model and threshold, new site)				U·D p	N	N	V·D+C	V·D+C c
A5	Locked external validation by evaluators other than the developers				N	N	N	N	N
A6	Temporal validation (later period, same source)				V·D p	N	N	N	N
A7	Operating-point performance at the declared threshold, external or temporal cohort				U·D p	N	N	U·D+C	U·D+C c
A8	Calibration, and predictive values at the intended prevalence				N	N	N	U·D	N c
A9	Performance by subgroup (age, sex, ancestry) at the declared threshold				N	N	N	U·D+C	N
A10	Device, manufacturer or signal-quality analysis				N	N	N	N	N
A11	Test-retest reliability				N	N	N c	N	N
A12	Equivalence across input forms, where the use admits more than one				N	U·D p	X	X	X
Level 3. Deployed intervention									
I1	Prospective silent evaluation in the target workflow				N	N	N	N	N

ID	Evidence requirement	ECGFounder (backbone)	ECG-FM (backbone)	DeepECG-SSL (backbone)	Hyperkalaemia, single lead (ECGFounder, fine-tuned)	AF detection, single lead (ECGFounder, frozen)	Structural heart disease (ECG-FM, adapted)	77-condition interpretation (DeepECG-SSL, fine-tuned)	Reduced LVEF (DeepECG-SSL, fine-tuned)
I2	Alert burden, predictive value in practice, yield per confirmatory test				N	N	N	N	N c
I3	Clinician response measured				N	N	N	N	N c
I4	Randomized or pragmatic trial, detection or process endpoint				N c	N	N	N	N c
I5	Randomized or pragmatic trial, patient-outcome endpoint				N	N	N	N	N
I6	Harms (false-positive burden, overdiagnosis) and equity reported				N	N	N	N	N
I7	Economic evaluation				N	N	N	N	N c

State: V, verified in the primary report; U, identified but unverified (abstract or secondary source, conflicting values or unconfirmed design); N, not identified in our search; X, inapplicable to the declared intended use. Evaluator after the dot: D, developers (for an adapted model, the group that built it, including groups independent of the backbone's developers); C, collaborating sites; I, independent of the evaluated model's developers; I?, independence uncertain (Supplementary Table S7; pairs classified I? are listed in Supplementary Table S10 and are not shown in a cell that independent (I) evaluations already establish). p, rests on a preprint. c, conventional task-specific evidence for the same clinical target exists (context only; not evidence about the model). Cells record whether evidence of a kind exists, not whether it is sufficient, and are not graded; N is not evidence that the model lacks the property or that no study exists. Backbone requirements are not scored for uses, and use requirements are not scored for backbones (blank). Consensus codes after independent coding by two authors and adjudication of disagreements (Supplementary Note 5), applied to a non-systematic search; the intended-use statements for the five uses are ours and illustrative. Backbone sources: ECGFounder [13], ECG-FM [16], DeepECG-SSL [6, 7]; independent evaluations [5, 11, 23, 24]; uses [6, 7, 15, 23, 24]. Operational definitions, eligibility rules and notes for every cell are in Supplementary Tables S6 to S11.

Table S6. State definitions and eligibility rules.

State	Meaning	Eligibility rule	How to read it
V	Verified	Evidence meeting the operational definition is reported; we read the full text of the primary report (journal article or preprint, including methods) and checked the item there; the values we use are consistent across the parts of the report we read. A preprint source adds the mark p.	A V cell says that evidence of this kind exists for this object. It does not say the evidence is sufficient, favourable or generalizable.
U	Identified, not sufficiently verified	Evidence that appears to meet the definition is reported, but at least one of the following holds: only an abstract, a secondary report or a supplementary summary was read; values differ between versions (preprint and journal) or between sections; or a design element required by the definition (for example patient-level separation) is not confirmed in the text read.	A U cell may become V or N after full-text reading. It must not be counted as verified evidence.
N	Not identified in our search	No evidence meeting the definition was identified in the sources we read, after the non-systematic search described in the review.	N is not evidence that the model lacks the property, that performance is poor, or that no relevant study exists anywhere. It records the limit of our search.
X	Not applicable to the declared intended use	The intended-use statement for the object excludes the condition the requirement tests, and the reason is recorded there (for example, a use declared for standard 12-lead input only does not need input-form equivalence; a backbone without text input has no prompt sensitivity).	X is never assigned because evidence is missing, and it cannot be assigned after the fact to avoid an N. Changing the intended use changes the X cells.
c (mark)	Conventional context	Evidence exists for a conventional, task-specific algorithm with the same clinical target.	Context only. The mark does not change the state of the cell and is not evidence about the foundation-model-based object.

Table S7. Evaluator classification.

Code	Meaning	Criteria
D	Developers	For this model-evaluation pair, at least one author was involved in developing the evaluated backbone or adapted model, or the report states that the evaluating group built it. For an adapted model, a group that built the adapted model is coded D even when it is independent of the backbone's developers (backbone-independent); data external to the developers do not make an evaluation independent.
C	Collaborating sites	Authors from sites that contributed data or evaluation to the developers' study, co-authoring with them.
I	Independent of the developers	For this model-evaluation pair: no author was involved in developing the evaluated model, and no institutional relationship or collaboration with its developers was identified in the author lists and affiliations read. Competing scientific or financial interests are recorded (declared, none declared, or no statement) and disclosed in Supplementary Table S10 but are not by themselves disqualifying; for example, evaluators who developed a competing model in the same comparison are coded I with that interest disclosed.
I?	Independence uncertain	Involvement in development, or an institutional relationship with the developers, could not be determined from the records read.
D+C, D+I	Combined	The cell rests on more than one evaluation, each classified separately for its own model-evaluation pair.

Table S8. Operational definitions and clinical justification of the 26 requirements.

ID	Requirement	Operational definition (what counts as evidence of this kind)	Clinical justification	Can it be X?
B1	Pretraining data separated from evaluation data (disjointness or overlap audit stated)	The report states that the evaluation data (patients or datasets) used to support the backbone's claims were excluded from pretraining, or reports an overlap audit.	Pretraining on evaluation patients inflates downstream estimates (leakage route, Section 4).	Not permitted.
B2	Reuse across two or more clinically distinct task families, with non-ECG reference standards	The backbone was adapted to at least two clinically distinct task families (for example rhythm, structure or function, laboratory value, outcome), at least one with a reference standard not derived from the ECG itself (echocardiography, laboratory value, registry outcome).	A platform claim requires reuse beyond labels that can be learned from ECG statements.	Not permitted.
B3	Label efficiency against a supervised comparator on the same labels	Performance at two or more reduced training-set sizes compared with a supervised model trained on the same labelled subsets, in the same population and test set.	Label efficiency is the main claimed advantage of pretraining (Section 3).	Not permitted.
B4	Backbone evaluated by investigators other than its developers	At least one evaluation of the backbone by investigators other than its developers; the evaluator code records whether independence was established.	Developer evaluations may favour the model through choices of task, tuning and comparator (Section 4).	Not permitted.
B5	Embedding analysis: source dataset versus label structure	Quantitative analysis of whether frozen embeddings cluster more by source dataset or acquisition than by diagnostic label (for example adjusted Rand index by dataset and by label).	Dataset-dominated geometry is a warning of possible shortcut use (principles P2 and P4).	Not permitted.
B6	Interventional test of reliance on acquisition features	An experiment that changes acquisition features (device, filtering, sampling rate, lead placement) while holding physiology fixed, with the effect on outputs measured.	Separates reliance on physiology from reliance on acquisition; embedding geometry alone cannot do this.	Not permitted.
B7	Prompt sensitivity of zero-shot use	For backbones with a text interface: variation of zero-shot outputs under paraphrased or reordered prompts.	Zero-shot outputs of ECG-language models can depend on wording (Section 7).	Permitted when the backbone takes no text input.
A1	Discrimination against a supervised comparator, same labels, target population	Discrimination of the adapted model compared with a model without pretraining (or the incumbent algorithm) trained on the same labelled data and evaluated on the same test set in the target population.	Without a comparator the contribution of the backbone, and the incremental value over existing tools, is unknown.	Not permitted.
A2	Evaluation patients stated to lie outside the pretraining corpus	The report states that evaluation patients, or their source, were not in the backbone's pretraining corpus.	Leakage from pretraining is invisible in the adapted model's own data split.	Not permitted.
A3	Patient-level separation of adaptation and test partitions stated	The report states that adaptation (training, tuning) and test partitions are disjoint at patient level.	Record-level splits place the same patient on both sides and inflate estimates.	Not permitted.
A4	Locked external validation (fixed model and threshold, new site)	Weights and threshold fixed before evaluation at a site or source not used for pretraining or adaptation.	Transport to a new site is the first test of deployability.	Not permitted.
A5	Locked external validation by evaluators other than the developers	A4 performed by evaluators other than the developers; the evaluator code records whether independence was established.	Independent replication is the strongest guard against selective reporting.	Not permitted.
A6	Temporal validation (later period, same source)	Evaluation of the locked model on a later period from the same source, excluding development patients.	Practice, populations and devices drift over time.	Not permitted.
A7	Operating-point performance at the declared threshold, external or temporal cohort	Sensitivity and specificity, or predictive values, at the declared, prespecified threshold, in an external or temporal cohort.	Clinical use happens at a threshold; AUROC does not fix alert burden (Sections 4 and 5).	Not permitted.
A8	Calibration, and predictive values at the intended prevalence	A calibration assessment (plot, slope and intercept, or equivalent), and predictive values at the intended prevalence, reported or derivable from reported counts.	Risk outputs and alert burden depend on calibration and prevalence.	Not permitted.

ID	Requirement	Operational definition (what counts as evidence of this kind)	Clinical justification	Can it be X?
A9	Performance by subgroup (age, sex, ancestry) at the declared threshold	Performance at the declared threshold reported for at least two of age, sex and ancestry or race-ethnicity groups.	Subgroup failure is a harm and an equity issue (Section 4).	Not permitted.
A10	Device, manufacturer or signal-quality analysis	Performance stratified by device, manufacturer or signal-quality level, or under controlled degradation.	Acquisition shift is a major cause of failure on transport.	Not permitted.
A11	Test-retest reliability	Agreement of outputs on repeated ECGs from the same patient within a clinically stable interval.	Serial and screening use require stable outputs.	Not permitted.
A12	Equivalence across input forms, where the use admits more than one	Where the intended use admits more than one input form (lead subsets, wearable or handheld versus 12-lead, sampling rates), performance shown for each form.	A model validated on one form cannot be assumed to work on another.	Permitted when the declared intended use specifies a single standard input form.
I1	Prospective silent evaluation in the target workflow	Prospective evaluation in the target workflow with outputs generated in real time but not shown to clinicians, compared with the reference standard.	Tests deployed behaviour (data flow, prevalence, alert rate) before patients are exposed.	Not permitted.
I2	Alert burden, predictive value in practice, yield per confirmatory test	Alerts per patient or per day, observed positive predictive value and yield per confirmatory test, measured in practice.	These quantities determine workload and the effect of an alert (Section 5).	Not permitted.
I3	Clinician response measured	Proportion of outputs acted on by clinicians, with reasons for not acting where possible.	The effect of an AI-ECG depends on a step the model does not control (Section 5).	Not permitted.
I4	Randomized or pragmatic trial, detection or process endpoint	Randomized trial (individual or cluster) of this model in use, with a detection or process endpoint.	Shows that use changes diagnoses or processes, with a counterfactual.	Not permitted.
I5	Randomized or pragmatic trial, patient-outcome endpoint	Randomized trial of this model in use with a patient-outcome endpoint (mortality, morbidity, events, patient-reported outcomes).	Only outcome endpoints show benefit to patients.	Not permitted.
I6	Harms (false-positive burden, overdiagnosis) and equity reported	Report of false-positive burden, downstream testing, overdiagnosis or adverse events, and of effects by population group.	Benefit has to be weighed against harm and its distribution.	Not permitted.
I7	Economic evaluation	Cost-effectiveness or budget-impact analysis using data from deployment of this model.	Health systems decide on value, not only on effect.	Not permitted.

Table S9. Declared intended uses of the worked examples and reasons for their selection.

Use	Declared intended use (ours)	Why it was selected and what it tests
Hyperkalaemia, single lead (ECGFounder, fine-tuned) [15]	Detection of hyperkalaemia, as defined in the source, from a single-lead (lead I) ECG in adults in hospital care, as a triage aid. Input form not fixed (a handheld prototype is described), so A12 is applicable.	Initialization transfer from a backbone with external evidence; the fullest adapted-model profile of the five; a non-structural laboratory target; a conventional randomized trial of a different potassium algorithm exists. Tests the limits of P1 and the scope of P7.
AF detection, single lead (ECGFounder, frozen) [23]	Detection of atrial fibrillation in short single-lead recordings. Input form not fixed in the source, so A12 is applicable.	The closest of the five to representation reuse (frozen encoder, separate classifier), evaluated by independent investigators. Tests the strongest case for P1 and the ceiling logic of P2.
Structural heart disease (ECG-FM, adapted) [24]	Multi-label screening for echocardiographic structural heart disease from a standard 12-lead ECG in adults; standard 12-lead input only (A12 not applicable).	Adaptation by independent investigators with an echocardiographic reference standard and a scratch comparator; no external validation. Tests P3 (support) and the value of comparators.
77-condition interpretation (DeepECG-SSL, fine-tuned) [6, 7]	Automated interpretation of 77 diagnostic labels from a standard 12-lead ECG; standard 12-lead input only (A12 not applicable).	A matched self-supervised versus supervised comparison from one group with locked external evaluation, showing parity. Shows that evidence for a use need not depend on an advantage of pretraining.

Use	Declared intended use (ours)	Why it was selected and what it tests
Reduced LVEF (DeepECG-SSL, fine-tuned) [6, 7]	Detection of left ventricular ejection fraction of 40% or less from a standard 12-lead ECG in adults; standard 12-lead input only (A12 not applicable).	The clinical target with the most complete conventional evidence chain, up to randomized trials and economic modelling (Section 5). Tests whether trial evidence can be bridged to a foundation-model successor (P6 and P7).

The intended-use statements are ours and are illustrative; they are not the developers' statements.

Table S10. Evaluator assessment by model-evaluation pair.

Evaluation	Evaluated model	Development involvement and institutional relationship	Competing scientific interests	Competing-interests statement	Code
ECGFounder development [13]	ECGFounder	Developers	Not applicable	In a supplement of the published article; not read	D
ECG-FM development [16]	ECG-FM	Developers	Not applicable	Not read	D
DeepECG preprint and article [6, 7]	DeepECG-SSL and DeepECG-SL	Developers, with co-authors from external centres	Not applicable	Read in the published article: two authors hold a pending patent on coronary angiogram analysis (unrelated); other authors list industry and research relationships	D+C
Hyperkalaemia preprint [15]	Model fine-tuned from ECGFounder	Senior author shared with ECGFounder	Not applicable	Not read	D
ICLR benchmark, Al-Masud et al. [5]	ECGFounder	Not involved in development; no institutional relationship identified (single university group)	Authors developed ECG-CPC and the S4 baseline in the same comparison (scientific interest, disclosed here)	None in the text read; funding: German Research Foundation	I
ICLR benchmark, Al-Masud et al. [5]	ECG-FM	As above	As above	As above	I
Holistic benchmark, Filice et al. [11]	ECGFounder; ECG-FM	Not involved in development; no relationship identified (two universities and a spin-off company)	None identified	None in the text read	I
FOUND-AF, Taleshinosrati et al. [23]	ECGFounder (frozen)	Not involved in development; no relationship identified (four universities)	Two authors co-authored cited non-foundation baselines	None in the text read	I
Structural heart disease adaptation, Do et al. [24]	ECG-FM	Not involved in development; no relationship with the ECG-FM developers (Toronto) identified	None identified	None in the text read	I
Structural heart disease adaptation, Do et al. [24]	ECGFounder	Not involved in development; some authors have Harvard-affiliated appointments, and overlap with ECGFounder co-authors' institutions was not assessed	None identified	None in the text read	I?
DeepECG benchmarking [6]	ECG-FM (14 overlapping labels, UK Biobank)	Not involved in development; no relationship with the ECG-FM developers identified	Developed competing models (DeepECG-SL, DeepECG-SSL) in the same comparison	Read (see DeepECG row)	I

Evaluation	Evaluated model	Development involvement and institutional relationship	Competing scientific interests	Competing-interests statement	Code
DeepECG benchmarking [6]	ECGFounder (47 overlapping labels, UK Biobank)	Not involved in development; external-centre co-authors include a Harvard-affiliated hospital, and overlap with ECGFounder co-authors' institutions was not assessed	Developed competing models in the same comparison	Read (see DeepECG row)	I?
ECG-CLIP development, Ko et al. [31]	DeepECG-SL and DeepECG-SSL; ECGFounder (comparators)	Not involved in development; no relationship identified (Scripps Research)	Developed a competing model (ECG-CLIP) in the same comparison	From the PubMed record: provisional patent held by four authors; advisory and consulting roles of one author. Primary text not read	I (abstract only)

Assessed separately for each model-evaluation pair from author lists, affiliations, funding and competing-interests statements in the texts we read (21 October 2026). Competing interests are disclosed, not treated as disqualifying. Financial relationships not stated in those texts could not be assessed. For adapted-model cells the evaluated model is the adapted model: the FOUND-AF authors and Do et al. are independent of the backbone developers (codes above) but built the adapted models they evaluated, so the corresponding cells for the AF and structural heart disease uses are coded D, backbone-independent (Supplementary Note 5).

Table S11. Notes and sources for every cell of the evidence matrix that is not a plain N.

Requirement / column	State (evaluator)	Basis for the entry	Sources
Level 1. Backbone			
B1 / BB1 ECGFounder	Identified, not verified (D), preprint	Consensus revision (previously N): CODE-test, PTB-XL and CinC2017 are labelled external evaluation sets, but the preprint text read states neither that they were excluded from pretraining nor an overlap audit; whether CODE is part of the pretraining corpus is unresolved, and the journal version was not read in full on this point.	[13]
B1 / BB2 ECG-FM	Verified (D)	UHN-ECG was excluded from pretraining and split by patient and time; the MIMIC-IV-ECG task shares a pretraining source. Main text read; supplementary tables not seen.	[16]
B1 / BB3 DeepECG-SSL	Verified (D+C)	Published Methods (full text read 21 Oct 2026): pretraining on development-hospital training ECGs (patient-level 70/10/20 split), CODE-15% and the MIMIC-IV training partition; MIMIC-IV was split 70/30 with patient partitioning, so MIMIC-IV-test is patient-disjoint from pretraining but from the same source. CLSA, PTB, UK Biobank and the seven external centres were not pretraining sources.	[6, 7]
B2 / BB1 ECGFounder	Verified (D+I)	Developers: rhythm, patient characteristics, NT-proBNP, ejection fraction and chronic kidney disease, mainly after fine-tuning. Independent benchmark: seven categories.	[5, 13]
B2 / BB2 ECG-FM	Verified (D+I)	Developers: UHN-ECG tasks and machine-read MIMIC-IV-ECG labels. Independent benchmark and an echocardiographic structural-disease study.	[5, 16, 24]
B2 / BB3 DeepECG-SSL	Verified (D)	Long QT detection and genotype classification, reduced ejection fraction (echocardiographic reference), 5-year incident atrial fibrillation and 77 report-derived labels (published full text). No independent evaluation of these uses found. Consensus revision: evaluator D, because collaborating-site data enter only the external evaluations, which this requirement does not need.	[6, 7]

Requirement / column	State (evaluator)	Basis for the entry	Sources
B3 / BB1 ECGFounder	Verified (I)	One controlled experiment on echocardiographic endpoints (three models, supervised S4 comparator, 250–1,000 examples): label-efficiency ratio 0.30–0.62 for ECGFounder (rechecked in the full text, 21 Oct 2026). Evaluators independent of the ECGFounder developers; they developed a competing model (ECG-CPC) and the S4 baseline in the same comparison (disclosed scientific interest), and the text gives no competing-interests statement.	[5]
B3 / BB2 ECG-FM	Verified (D)	Developers' learning curves: linear probe best at the smallest sizes; baselines converge as data grow. The independent label-efficiency experiment did not include ECG-FM.	[5, 16]
B3 / BB3 DeepECG-SSL	Identified, not verified (D)	Published text states that the SSL-SL gap widened as training data decreased and reports models trained from random initialization; its long QT comparison (pretrained 0.875 vs random initialization 0.602) conflicts with Supplementary Table S37 of the DeepECG article, which labels both values as randomly initialized models for long QT detection. Supplementary Table S37 of the DeepECG article (read 21 Oct 2026) gives the random-initialization values (for LVEF <50% the supervised model trained from random initialization, 0.909, exceeded both fine-tuned models, 0.885 and 0.887); the learning curves by training-set size are in Supplementary Figure 4, an image that is not in the supplementary files read. The supervised comparator had a 670-run Bayesian sweep; equivalent tuning of the self-supervised model is not described. Consensus revision: evaluator D; the evidence is the developers' report, and the possible independent source (Ko et al.) was not read.	[6, 7]
B4 / BB1 ECGFounder	Verified (I)	Benchmark (ICLR 2026) and two preprints by groups not involved in developing ECGFounder and with no institutional relationship identified (I); a third preprint has Harvard-affiliated authors, and institutional overlap with ECGFounder co-authors was not assessed (I? for that pair). Developers of competing models also benchmarked ECG-Founder: DeepECG on 47 overlapping labels in UK Biobank (full text read; external-centre co-authors include a Harvard-affiliated hospital, so I? for that pair) and ECG-CLIP as a comparator (abstract read). No competing-interests statements in the four external texts read. A further peer-reviewed benchmark identified during revision, BenchECG [25], also evaluated ECGFounder (authors developed a competing model and declared no competing interests); state unchanged (Supplementary Note 4).	[5,6,11,23,24,31]
B4 / BB2 ECG-FM	Verified (I)	Benchmark (ICLR 2026) and two preprints by groups not involved in developing ECG-FM and with no institutional relationship identified; no competing-interests statements in the texts read. DeepECG developers (a competing model) also compared ECG-FM on 14 overlapping labels in UK Biobank (published full text read; I, with disclosed competing interest).	[5, 6, 11, 24]
B4 / BB3 DeepECG-SSL	Identified, not verified (I)	ECG-CLIP developers (Scripps; no relationship with the DeepECG developers identified) used DeepECG-SL and DeepECG-SSL as comparators, with external evaluation on MIMIC-IV; abstract read only, so U. The evaluators developed a competing model (disclosed interest), and DeepECG-SSL pretraining included a MIMIC-IV training partition. Not among the models in the four independent benchmarks.	[5,6,11,23,31]
B5 / BB1 ECGFounder	Verified (I), preprint	Frozen embeddings: adjusted Rand index by dataset 0.03 (conduction) and 0.12 (AF); by label 0.24 and 0.70.	[11]
B5 / BB2 ECG-FM	Verified (I), preprint	Frozen embeddings: adjusted Rand index by dataset 0.31 (conduction) and 0.20 (AF); by label 0.23 and 0.64.	[11]

Requirement / column	State (evaluator)	Basis for the entry	Sources
B6 / BB1 ECGFounder	Identified, not verified (I), preprint	Consensus revision (previously N): a preprint identified by the second rater, Nanayakkara and Meegama, listed as 'Beyond in-distribution benchmarks: a controlled evaluation of foundation and from-scratch models for 12-lead ECG quality assessment' (SSRN 6753359, 2026), was assessed from its public abstract only; the full text could not be accessed. The abstract reports a comparison of ECGFounder, ECG-FM, HuBERT-ECG, ResNet-1D and InceptionTime in seven evaluation settings, including synthetic cross-dataset generalization and artefact-specific robustness with noise injected into Chapman-Shaoxing recordings, and gives ECGFounder-specific results (AUROC 0.829 at 5 dB; strong robustness to muscle artefact and electrode motion). Because the task is ECG quality assessment, whether the perturbations change acquisition features with physiology held fixed, and how they relate to the quality labels being predicted, cannot be determined from the abstract. A qualifying B6 intervention is therefore not confirmed and the cell is not counted as verified evidence; it becomes V or N after full-text reading against the B6 definition.	SSRN 6753359 (full text not read; not in the reference list)
B6 / BB2 ECG-FM	Identified, not verified (I), preprint	As for B6 / BB1, for ECG-FM. The abstract lists ECG-FM among the compared models but gives no ECG-FM-specific result, so the candidate evidence is less specific for this backbone than for ECGFounder.	SSRN 6753359 (full text not read; not in the reference list)
B7 / BB1 ECGFounder	Inapplicable	The backbone takes no text prompt. The requirement applies to ECG-language models (Section 7).	
B7 / BB2 ECG-FM	Inapplicable	The backbone takes no text prompt. The requirement applies to ECG-language models (Section 7).	
B7 / BB3 DeepECG-SSL	Inapplicable	The backbone takes no text prompt. The requirement applies to ECG-language models (Section 7).	
Level 2. Adapted model			
A1 / U3 Structural heart disease	Verified (D), preprint	Macro AUROC on the test split: scratch ResNet1D 0.8246; ECG-FM probe after in-domain adaptation 0.8233; partial fine-tuning 0.8509. Comparators protocol-aligned, not parameter-matched; scratch-model tuning not described; single point estimates. Consensus revision: evaluator D, because the authors built the adapted model; they are independent of the ECG-FM developers (backbone-independent).	[24]
A1 / U4 77-condition interpretation	Identified, not verified (D+C)	Micro-AUROC, self-supervised versus supervised: 0.990 vs 0.992 internal; 0.981 vs 0.980 on four external public datasets; 0.983 vs 0.983 on seven external private datasets (published full text read 21 Oct 2026; consistent with the preprint). Tuning was not matched (670-run sweep for the supervised model). Consensus revision: U, because the internal self-supervised value differs between sections and versions (0.989 in the preprint's Results; 0.990 in its Abstract and in the journal article).	[6, 7]
A1 / U5 Reduced LVEF	Identified, not verified (D+C)	Supplementary Table S25 of the DeepECG article (read 21 Oct 2026): internal test set (25,252 ECGs; 2,093 with LVEF 40% or lower) AUROC 0.926 (0.923–0.929) self-supervised vs 0.917 (0.913–0.921) supervised, matching the Results. The structured summary gives 0.928 vs 0.900, which is not found in the tables read and is unexplained. Tuning was not matched (670-run sweep for the supervised model). For LVEF below 50% the supervised model trained from random initialization (0.909, DeepECG article Table S37) exceeded both fine-tuned models (0.885 and 0.887). Consensus revision: U, because the values in the structured summary and the preprint (0.928 vs 0.900) differ from those in the Results and Table S25, and the supervised comparator was initialized from the supervised interpretation model rather than trained without pretraining.	[6, 7]
A2 / U1 Hyperkalaemia, single lead	Identified, not verified (D), preprint	Consensus revision (previously N): the report states that the backbone was pretrained on a multinational corpus of more than ten million recordings but not whether the development and external hospitals were in it; separation could only be inferred by combining reports.	[13,15]

Requirement / column	State (evaluator)	Basis for the entry	Sources
A2 / U2 AF detection, single lead	Verified (D), preprint	The authors (not involved in developing ECGFounder; no institutional relationship identified; no competing-interests statement) state that none of the four datasets was explicitly included in pretraining and that complete exclusion of overlap cannot be guaranteed. Consensus revision: evaluator D, because the authors built the adapted pipeline; they are independent of the ECGFounder developers (backbone-independent).	[23]
A2 / U3 Structural heart disease	Identified, not verified (D), preprint	Consensus revision (previously N): the ECG-FM report lists its pretraining datasets, which do not include EchoNext (Columbia), but the evaluating report does not state this, and its own continued self-supervised adaptation used EchoNext training waveforms. The authors are independent of the ECG-FM developers (backbone-independent).	[16,24]
A2 / U4 77-condition interpretation	Verified (D+C)	Published Methods: the private external centres, CLSA, PTB and UK Biobank were not pretraining sources; MIMIC-IV-test is patient-disjoint from the MIMIC-IV training partition used for pretraining but comes from the same source.	[6]
A2 / U5 Reduced LVEF	Verified (D+C)	LVEF models fine-tuned on the development training split and tested on its patient-level test split and on CSH, NYP, UCSF, UW, JGH and MIMIC-IV (published Methods); only the training split and the MIMIC-IV training partition entered pretraining.	[6]
A3 / U1 Hyperkalaemia, single lead	Verified (D), preprint	Development patients split 8:1:1 at patient level; the temporal set excluded development-period patients; the external hospital was reserved for validation.	[15]
A3 / U2 AF detection, single lead	Identified, not verified (D), preprint	Consensus revision (previously N): five-fold cross-validation grouped by recording identity; the contributions mention patient- or recording-level grouping, and whether patients contribute several recordings is not stated. The authors built the adapted pipeline (backbone-independent).	[23]
A3 / U3 Structural heart disease	Identified, not verified (D), preprint	Release-defined splits were used; the authors state they created no new patient-level or record-level splits; validation and test keep each patient's latest ECG; patient-disjointness is not explicitly confirmed. Consensus revision: evaluator D, backbone-independent, as for A1.	[24]
A3 / U4 77-condition interpretation	Verified (D)	Development-hospital data split randomly 70/10/20 at patient level, with no patient in more than one split (published Methods).	[6]
A3 / U5 Reduced LVEF	Verified (D)	LVEF tasks use the patient-level development split (training vs test); each ECG paired with the closest echocardiogram from -90 to +7 days, one per patient (published Methods).	[6]
A4 / U1 Hyperkalaemia, single lead	Identified, not verified (D), preprint	Prespecified frozen threshold (value not reported) applied to an external hospital's 2024 ECGs: AUROC 0.808 (95% CI 0.788–0.826). The senior author is also senior author of ECGFounder. Consensus revision: U, because the report does not state that the internally prespecified threshold was applied unchanged at the external hospital, or whether that hospital's ECGs were in the backbone's pretraining corpus.	[13, 15]
A4 / U4 77-condition interpretation	Verified (D+C)	Per-label thresholds optimized on the development-hospital validation split and applied unchanged to all cohorts (published Methods); investigators of the private centres are co-authors.	[6]
A4 / U5 Reduced LVEF	Verified (D+C)	Published Methods: thresholds from the development validation split applied to six external cohorts. Supplementary Tables S26-S33 of the DeepECG article (read 21 Oct 2026) report external AUROC for LVEF 40% or lower (self-supervised 0.881–0.907, supervised 0.853–0.888, six cohorts and their pooled total); investigators of the external centres are co-authors. Conventional context: a four-site external validation of a commercial algorithm, 13,960 patients, AUROC 0.92 (0.91–0.93), by Mayo Clinic and other-site investigators and company employees.	[6,35]
A6 / U1 Hyperkalaemia, single lead	Verified (D), preprint	Same-hospital ECGs after July 2021: AUROC 0.858 (0.844–0.872).	[15]

Requirement / column	State (evaluator)	Basis for the entry	Sources
A7 / U1 Hyperkalaemia, single lead	Identified, not verified (D), preprint	Temporal sensitivity 85.05% (95% CI 81.84–87.77%); external negative predictive value above 99.3%. External sensitivity, specificity and positive predictive value at the primary threshold are not reported. Consensus revision: U, because threshold transport is not stated (see A4) and only the negative predictive value is reported externally at the primary threshold.	[15]
A7 / U4 77-condition interpretation	Identified, not verified (D+C)	Supplementary Tables S11-S24 of the DeepECG article (read 21 Oct 2026) report overall and per-label sensitivity and specificity for the internal test set and each external cohort, and list per-label thresholds in two columns (PPV-NPV and sensitivity-specificity). The tables do not state which threshold produced the reported values, and the self-supervised model's reported operating point (internal overall sensitivity 0.994, specificity 0.795, against 0.959 and 0.952 for the supervised model) is not the balanced point described in the Methods, so performance at the declared threshold is not directly verified. The second rater coded V; consensus retained U (Table S17).	[6]
A7 / U5 Reduced LVEF	Identified, not verified (D+C)	Supplementary Tables S25-S33 of the DeepECG article (read 21 Oct 2026) report sensitivity and specificity for LVEF 40% and 50% cut-offs internally and externally, but thresholds are listed only for the internal test set (sensitivity-specificity threshold 0.002 self-supervised, 0.197 supervised), and the tables do not state which threshold produced the reported values. The self-supervised operating point is extreme (LVEF 40% or lower: internal sensitivity 0.993, specificity 0.454; external sensitivity 0.958–0.988, specificity 0.276–0.435), so performance at the declared threshold is not directly verified. Conventional context: sensitivity 84.5% and specificity 83.6% for a commercial algorithm in a four-site validation. The second rater coded V; consensus retained U (Table S17).	[6,35]
A8 / U4 77-condition interpretation	Identified, not verified (D)	Supplementary Table S43 of the DeepECG article (read 21 Oct 2026) gives the weighted Brier score by dataset and diagnostic category: lower with self-supervised pretraining by 0.013 on PTB and MIMIC-IV and 0.006 on UK Biobank, equal on CLSA and higher on the internal test set (0.078 vs 0.072); category-level changes are mixed. The published article presents this weighted Brier score as its calibration evaluation (Methods). Table S8 requires a calibration assessment (plot, slope and intercept, or equivalent) and predictive values at the intended prevalence, reported or derivable from reported counts; whether the weighted Brier score is an equivalent assessment is not confirmed, and no predictive values at an intended prevalence are reported (the PPV-NPV columns of the performance tables are thresholds), nor was derivability from reported counts established. Coded U under Table S6 (a required element not confirmed in the text read); not counted as verified calibration evidence. Consensus revision: evaluator D; the evidence uses public datasets and the internal test set, without collaborating-site data.	[6]
A8 / U5 Reduced LVEF	Not identified	Conventional context: positive predictive value 30.5% and negative predictive value 98.4% at a prevalence of 7.9%.	[35]
A9 / U4 77-condition interpretation	Identified, not verified (D+C)	Published Methods: equalized odds at the optimized thresholds across three age groups and sex. The main text reports only pooled bounds (true-positive-rate differences below 0.1, false-positive-rate differences below 0.02, micro-averaged across the 77 conditions; below 0.01 in the preprint). The DeepECG article's supplement presents the differences in its Supplementary Figure 14 (panels for sex and age in five datasets), an image that is not in the supplementary files read; no table of group-specific rates was found, so performance by group at the declared threshold is not directly verified. A pooled bound does not limit disparity for any single condition; ancestry not reported. The second rater coded V; consensus retained U (Table S17).	[6, 7]
A11 / U3 Structural heart disease	Not identified	Conventional context: a test-retest preprint of single-target networks and an ensemble for structural heart disease (developer group; different models and labels).	[19]

Requirement / column	State (evaluator)	Basis for the entry	Sources
A12 / U2 AF detection, single lead	Identified, not verified (D), pre-print	Consensus revision (previously N): performance is reported for each input form (native short single lead at 300 Hz; Holter lead II at 128 to 250 Hz), but each result comes from a classifier fitted and evaluated within that form's dataset, so it is not confirmed that the reported performance is that of one adapted model applied to each form, which is what the requirement addresses (Table S8: a model validated on one form cannot be assumed to work on another). Coded U under Table S6 (a required element not confirmed in the text read); not counted as verified evidence of equivalence.	[23]
A12 / U3 Structural heart disease	Inapplicable	Declared input: standard 12-lead ECG; the use does not admit other forms.	
A12 / U4 77-condition interpretation	Inapplicable	Declared input: standard 12-lead ECG; the use does not admit other forms.	
A12 / U5 Reduced LVEF	Inapplicable	Declared input: standard 12-lead ECG; the use does not admit other forms.	
Level 3. Deployed intervention			
I2 / U5 Reduced LVEF	Not identified	Conventional context (EAGLE): about 6% of ECGs positive; echocardiography in 49.6% vs 38.1% of AI-positive patients; among those tested, EF 40% or less in 22.1% (intervention) vs 22.7% (control) as reported in the cost-effectiveness inputs (Section 5).	[39,41]
I3 / U5 Reduced LVEF	Not identified	Conventional context (EAGLE secondary analysis): echocardiography after 78.7% vs 43.0% of AI-positive ECGs for high vs other adopters; a post hoc analysis of a liver-disease trial found the same dependence on clinician engagement (Section 5).	[40,50]
I4 / U1 Hyperkalaemia, single lead	Not identified	Conventional context: a physician-randomized trial of an AI-ECG potassium alert in emergency patients (abstract read; a different algorithm): no difference in treatment overall, more treatment among AI-identified hyperkalaemia (Section 5).	[45]
I4 / U5 Reduced LVEF	Not identified	Conventional context: EAGLE (22,641 patients; new EF 50% or less in 2.1% vs 1.6%), an inpatient trial (13,631) and an obstetric trial (1,232); EF thresholds and algorithms differ from this model's.	[39,42,46]
I7 / U5 Reduced LVEF	Not identified	Conventional context: modelled cost-effectiveness of EAGLE, $27,858 per quality-adjusted life-year.	[41]

Cells scored N without a conventional-context mark are not listed; for them the entry means only that we did not find evidence of the kind in the records we read, and the search was not systematic. Entries marked "Conventional context" concern a different algorithm for the same clinical target and are not evidence about the model in the column. Bibliographic status of each source is in the master RIS library. Rows marked 'Consensus revision' changed at adjudication (Supplementary Note 5 and Table S17).

Table S12. Trial-registry search (ClinicalTrials.gov API v2 and ISRCTN, 21 October 2026).

Registry; query	Records	ECG-relevant records and assessment
ClinicalTrials.gov: "foundation model" AND (ECG OR electrocardiogram OR electrocardiography)	2	NCT07395648 [55]: RCT of smartwatch with unspecified pre-trained foundation model, post-MI, not yet recruiting. NCT06591923: observational development of a multimodal foundational model (ECG, imaging, laboratory), status unknown
ClinicalTrials.gov: "foundation model" / "foundational model" / "foundation models" (all records screened for any ECG mention in the full protocol)	27 / 5	Only the two records above mention ECG; others concern pathology, imaging, EHR or education
ClinicalTrials.gov: "self-supervised" AND (ECG OR electrocardiogram)	0	None

Registry; query	Records	ECG-relevant records and assessment
ClinicalTrials.gov: "pretrained" AND ECG AND (AI OR deep learning)	0	None
ClinicalTrials.gov: "ECG-Founder"; "ECG-FM"; "HuBERT-ECG"	0; 0; 0	None
ClinicalTrials.gov: "DeepECG"	3	HEART-AI NCT06462989 [56] (RCT, EchoNext score via DeepECG.ai, frozen weights, enrolling); DAISEA-ECG NCT06637293 [57] (stepped-wedge, DeepECG platform, model not specified, enrolling); NCT07636759 (observational; a different Korean algorithm named DeepECG LVSD)
ClinicalTrials.gov: "large language model" AND electrocardiogram	3	NCT07810686 ECG-IA [58]: RCT of version-locked GPT-4o assistance for prehospital ECG vignettes, recruiting
ISRCTN: "foundation model"; "self-supervised" ECG; ECG "large language model"; ECG "deep learning"; electrocardiogram "artificial intelligence"	1; 0; 0; 0; 4	None tests an ECG foundation model (ISRCTN13770753 pathology; ISRCTN39127320, 13724416, 17993837, 55166064 other designs)

Classification of the relevant interventional records (from the registered intervention descriptions, 21 October 2026): (1) may evaluate an ECG foundation model directly: NCT07395648 (model identity not stated); (2) conventional or unspecified algorithms from an ECG foundation-model developer group: NCT06462989 (task-specific EchoNext score, frozen weights) and NCT06637293 (model not specified); (3) general-purpose multimodal language model: NCT07810686 (GPT-4o). NCT06591923 and NCT07636759 are observational. WHO ICTRP, ChiCTR, EU CTIS, ANZCTR and other national registries were not searched; claims about the absence of trials are limited to the registries listed.

Supplementary Note 4. Analyses added during revision

Revision search (5 November 2026). To address a reviewer's request for broader comparative synthesis, we ran a targeted web search on 5 November 2026 for multi-model benchmarks of ECG foundation models (terms: benchmark, ECG foundation models, independent evaluation). This search is separate from the search described in Supplementary Note 1, was not systematic and covered web-indexed articles and preprints only. It identified one additional comparison of several published foundation models by investigators who had not developed them: BenchECG [25], first posted as a preprint in September 2025 and published in npj Digital Medicine on 14 September 2026 (accelerated article preview; full text and supplementary information read on 5 November 2026; values below are from the published version, which differs from the preprint in the models compared and in several scores). Its authors developed one compared model, xECG, and declared no competing interests. They evaluated the released weights of ST-MEM, ECG-JEPA, ECG-CPC and ECGFounder alongside xECG on eight public datasets (421,171 patients; 1,674,704 recordings) and ten tasks, under fine-tuning and linear probing with five runs per task. Fine-tuned BenchECG scores were 0.838 for xECG, 0.797 for ECG-CPC (2.9 million parameters), 0.793 for ST-MEM (85.2 million), 0.783 for ECG-JEPA (85.4 million) and 0.588 for ECGFounder (30.7 million), and 0.769 for a supervised xLSTM trained without pretraining. Rankings depended on the adaptation regime: for example, ECG-JEPA's AUROC for atrial fibrillation from photoplethysmography was 0.477 with linear probing and 0.818 after fine-tuning.

Out-of-distribution testing in BenchECG. BenchECG labels four task settings as out-of-distribution. For age estimation and mortality prediction, models were fine-tuned on CODE-15% (Brazil, primary care; 5-year mortality 4.7%) and

applied unchanged to MIMIC-IV-ECG (USA, intensive care; 5-year mortality 36%) and, for age, to PTB-XL (Germany) and CPSC2018 (China); none of these test datasets was among the pretraining datasets that the benchmark authors list for the compared models, whereas CODE-15% was in the pretraining data of three models and was used only for fine-tuning. Fine-tuned foundation models reached a C-index of 0.677 to 0.710 on MIMIC-IV-ECG (0.626 for the supervised xLSTM), with drops from the CODE-15% validation set of 0.082 to 0.133 (0.024 for the supervised xLSTM); mean test mean absolute errors for age were 7.560 to 8.854 years (8.896 for the supervised xLSTM), and the validation-to-test increase was 1.49 years for xECG and, by our arithmetic, 0.92 years for the supervised xLSTM. The other two settings, atrial fibrillation from photoplethysmography and R-peak detection in exercise ECGs, were fine-tuned on data from the target setting and test adaptation to a new signal rather than transport of a fixed model. We classify the age and mortality tasks as independent tests of cross-population transport of fixed, fine-tuned models for four of the five models (the authors developed xECG). They are not locked external validation of a clinical use in the sense of Table 2: the targets are a patient characteristic and a ranking of mortality risk, the metrics are threshold-free, no operating threshold or calibration was assessed, and the authors state that clinical adoption would require prospective validation and assessment of calibration. The matrix codes are unchanged, because BenchECG does not evaluate ECG-FM, DeepECG-SSL or any of the five worked uses, and the ECGFounder backbone cells it bears on (B2, B4) were already verified.

Other records identified. The revision search also identified a multi-task benchmark comparing ECG-FM with general time-series and language models on MIMIC-IV-ECG [59], in which ECG-specific pretraining did not clearly outperform general time-series pretraining (one author developed one of the time-series models compared), and benchmarks of self-supervised methods trained by their authors [60] or of interpretability [61], which do not compare published foundation models on clinical performance.

Reproducibility check of the matrix tallies (not a validation). Both raters' original codes and the consensus codes are given in machine-readable form in Supplementary Data 2. The first rater's codes were checked cell by cell against Table S5 as submitted with the original manuscript (116 of 116 identical), and the tallies reported in Section 6.5 were recomputed from the consensus codes: backbone cells 11 V, 5 U, 2 N and 3 X; use cells 10 V, 13 U, 69 N and 3 X; 23 identified use-level cells, all at the adapted-model level and all from the groups that built the adapted model (9 with collaborating sites; 6 from groups independent of the backbone's developers); and 35 intervention cells, none identified. This establishes that the reported counts follow from the codes, not that the codes are correct. Agreement between the raters is reported separately (Supplementary Note 5).

Exploratory intended-use sensitivity analysis. Table S8 permits the state X for two requirements only: B7 (prompt sensitivity), when a backbone takes no text input, and A12 (equivalence across input forms), when a use declares a single standard input form. The declared intended use can therefore change at most 8 of the 116 cells through applicability (3 B7 cells and 5 A12 cells), of which 6 are currently X. Table S14 re-applies the documented eligibility rules to the consensus codes of Table S5 under alternative declared uses; no code was changed. A cell that becomes applicable is given the state supported by the sources already read, which in every case is N, and a cell that becomes

inapplicable is X whatever its state (in S2 this removes one unverified cell); a fresh search under the changed use could identify evidence we did not seek. Scenario S2 quantifies the effect of narrowing a use, which the rule that X cannot be assigned after the fact to avoid an N is intended to prevent, and is not endorsed. Scenario S5 shows that the declared population, through the support principle (P3), can change verified and unverified cells, not only cells marked inapplicable. The analysis is exploratory: it tests how much the matrix depends on the authors' applicability choices, not whether the framework or its codes are valid. Results were recomputed from the consensus codes by a script and by the spreadsheet formulas of Supplementary Data 2, with identical results.

Table S13. Provisional decision rules for classifying a model as an ECG foundation model, with examples from the sources reviewed (our proposal; untested for agreement between assessors).

Criterion	Provisional rule	Counts	Does not count	Borderline cases (record the rationale)
Broad pretraining data	Pretraining data not selected on any downstream target and drawn from more than one source or from an unselected clinical archive. No numerical threshold for size.	Public multi-source collections (ST-MEM [28]); institutional archives not selected on a target (ECG-Founder [13]; paediatric archive of ECG-Fyler [8])	A cohort assembled for one target and then reused as pretraining data	A large single-institution archive limited to one care setting
Clinically distinct task	Target defined by a different clinical question or reference standard from the pretraining targets. Supervised: neither a pretraining label nor a merger or relabelling of pretraining labels. Self-supervised: any clinical target. Language-aligned: reference standard outside the paired reports.	ECGFounder: natriuretic peptide and ejection fraction [13]; ST-MEM: diagnostic classes beyond reconstruction [28]	ECGFounder evaluated on PTB-XL classes that overlap its 150 training labels; the same label set in a new dataset (tests transfer across populations or acquisition, not a new task)	A clinical composite of pretraining labels; a language-aligned model evaluated on diagnoses that appear in its reports
Data not used for pretraining	Exclusion of evaluation data from pretraining documented at patient or source level.	ECG-Fyler excluded test patients from pretraining [8]; UHN-ECG excluded from ECG-FM pretraining [16]	Evaluation on a dataset that was a pretraining source without a patient-level statement	Same source, patient-disjoint partitions (DeepECG MIMIC-IV test [6]): counts, with the shared source recorded
Pretrained on ECGs	Representation learned from ECG data, alone or with other modalities.	All backbones in the evidence matrix	A network initialized from ImageNet weights and fine-tuned on ECG-derived images [62]	Models pretrained on photoplethysmography or other biosignals and adapted to ECG
Classification versus evidence	Classification needs one qualifying transfer and does not depend on the evaluator; the strength of a reuse claim is described separately by the reuse evidence profile (Section 2.6).	-	-	-

The rules operationalize the criterion of Section 2.6. They are provisional, set no numerical thresholds and have not been tested for agreement between assessors; they were not part of the agreement study (Supplementary Note 5), which covered the matrix codes only. Classification as a foundation model says nothing about the strength of evidence for any use.

Table S14. Exploratory intended-use sensitivity analysis of the evidence matrix (consensus codes; no code changed).

Scenario	Alternative declared use or assumption	Cells changed	Backbone cells (V/U/N/X)	Use cells (V/U/N/X)	Applicable cells	Identified cells (V or U)	Effect on interpretation
S0	Base case (declared uses as in Table S9)	0	11/5/2/3	10/13/69/3	110	39	Consensus matrix (Table S5).
S1	Uses U3, U4, U5 extended to reduced-lead, wearable or image input	3	11/5/2/3	10/13/72/0	113	39	A12 becomes applicable to three 12-lead uses; no input-form evidence identified, so X to N. No identified evidence added; conclusions unchanged.
S2	U1 and U2 restricted to a single, fixed input form (lead I at the source sampling rate)	2	11/5/2/3	10/12/68/5	108	38	A12 excluded for both single-lead uses: N to X for U1 and U to X for U2. Narrowing removes one not-identified and one unverified cell from view; this is what the rule against after-the-fact X guards against.
S3	Backbones treated as having an ECG-text (zero-shot) interface	3	11/5/5/0	10/13/69/3	113	39	Counterfactual: B7 applicable to the three backbones, X to N. Signal-only backbones cannot have prompt-sensitivity evidence.
S4	Most permissive: S1 + S3 (no X cells remain)	6	11/5/5/0	10/13/72/0	116	39	Upper bound: no X cells remain; 6 cells change; identified evidence unchanged (39 cells).
S5	Population scope: U3 and U5 extended from adults to children (support rule, P3)	2	11/5/2/3	9/12/71/3	110	37	Not an X change: A1 for U3 (verified) and U5 (unverified), established in adults, becomes N for a paediatric stratum (P3, support). Declared population matters more than applicability.

Scenarios re-apply the eligibility rules of Tables S6 and S8 and the support principle (P3) to the codes of Table S5. Proportions are descriptive, not scores. Calculations are reproduced in Supplementary Data 2.

Table S15. Proposed minimum documentation of pretraining–evaluation separation for ECG foundation-model studies (our proposal, not an existing standard).

Item	What to report	Why	Matrix requirement
1. Pretraining corpus provenance	Datasets, versions, institutions, calendar periods, numbers of patients and recordings; or, if the corpus cannot be released, hashed record or patient identifiers	Third parties cannot judge disjointness from an undocumented corpus	B1
2. Source-level overlap	For each evaluation dataset, whether its source entered pretraining, adaptation or threshold selection	A dataset external to a hospital may be internal to the pretraining corpus	B1, A2
3. Patient- and record-level overlap	Method used to check overlap (identifier matching; otherwise near-duplicate signal detection) and its result	Patients contribute many similar ECGs; individual identity is recoverable from the signal	A2, A3
4. Patient-level partitioning	All of a patient's recordings in one partition of the adaptation data	Record-level splits inflate estimates	A3
5. Temporal separation	Calendar periods of pretraining, adaptation and evaluation data	Drift in practice, devices and case mix	A6
6. Labels and reference standards	Origin of labels in each phase; whether evaluation labels derive from machine interpretation or from readers who saw the tracing	Agreement with a reading is not agreement with the patient's condition	A1, A8
7. Tuning and threshold selection	Data used to choose hyperparameters, checkpoints and thresholds	Thresholds chosen on test data are not locked	A4, A7

Item	What to report	Why	Matrix requirement
8. Three attributes reported separately	Institutional externality, pretraining disjointness and evaluator independence (Table S7)	Each can hold without the others	A4, A5, B4

Our proposal, integrated with the evidence matrix (requirement codes as in Table S8); it is not an existing reporting standard and has not been tested.

Supplementary Note 5. Independent double coding, agreement and adjudication of the evidence matrix

Procedure followed. The first rater (M.A.M.) coded all 116 cells for the original submission, with the AI-assisted extraction described in Supplementary Note 1; his codes are those of Table S5 as originally submitted, and the copy used for comparison was transcribed from that table, not taken from a separately time-stamped coding file. The second rater (R.Q.), an author who had not taken part in the original coding, coded all 116 cells during revision using the coding form of Supplementary Data 1, the coding guide (Tables S6 to S8), the declared uses (Table S9) and the primary sources, without access to the first rater's codes, Figure 5 or Tables S5 and S11, as reported by both raters. Both raters subsequently confirmed their respective assessments. M.A.M. confirmed that the original Table S5 classifications faithfully represented his first assessment, and R.Q. confirmed that he personally reviewed and approved all 116 classifications in his independent second assessment, including those initially supported by AI-assisted evidence extraction. Their original classifications were preserved unchanged for the pre-adjudication agreement analysis.

Deviations from the plan. The plan in Supplementary Data 1, retained unchanged (Supplementary Note 1), specified two raters not involved in the Review, joint training on 26 cells (the ECG-FM backbone and the structural heart disease use) excluded from analysis, and adjudication by a third rater. Before the second coding the authors instead chose to code all 116 cells, with a fictitious worked example for orientation; no training on real cells took place, the second rater was an author, and the first rater's codes were those recorded in Table S5 of the original submission rather than a newly completed coding exercise. We therefore report all 116 cells as the principal analysis and the 90 cells the plan would have analysed as a sensitivity analysis. The plan was not registered, and neither analysis was prespecified beyond Supplementary Data 1. Its decision rule, to redefine any requirement with AC1 below 0.6, could not be applied as intended because each requirement has only three to five cells; disagreements concentrated in B6 (2 of 3 cells), A7 (3 of 5), A1 and A2 (2 of 5 each), and the coding guide was not revised, the adjudication instead applying its existing rules strictly (Table S17).

Agreement before adjudication. Agreement was computed from the two raters' original state codes before any discussion or adjudication, never from consensus codes (Table S16). The raters agreed on 101 of 116 cells (87.1%; Cohen's κ 0.759; Gwet's AC1 0.843) and on 77 of the 90 cells of the planned analysis set (85.6%; κ 0.728; AC1 0.825). The first rater coded 25 cells V, 7 U, 78 N and 6 X, the second 26 V, 14 U, 70 N and 6 X; the second rater used U more often. Of the 15 disagreements, 6 were N against U, 4 V against U, 3 U against V and 2 N against V; none involved X. In the 35 intervention cells both raters used only N, so κ is undefined there (expected agreement is 1) and AC1, computed over the four categories of the scale, equals 1 trivially. Because N dominates, κ is

sensitive to the marginal distribution, which is why AC1 is also given; neither statistic was used to set a threshold for acceptable agreement.

Uncertainty. Cells are not independent: they are nested in 8 objects and 26 requirements, and many share sources, so we report no confidence intervals, and the plan's bootstrap over cells, which would treat them as independent, was not used. For orientation only, cluster bootstrap resampling of objects (8 clusters, 5,000 resamples) gave a percentile range of 0.66 to 0.85 for κ, and resampling of requirements (26 clusters) 0.64 to 0.87; with so few clusters these ranges are unstable and are not 95% confidence intervals. Omitting one object at a time gave κ from 0.73 to 0.80 and AC1 from 0.82 to 0.86.

Evaluator codes and marks. Of the 25 cells that both raters coded with the same identified state (V or U), they gave the same evaluator code in 14 (56%). In the other 11 the second rater classified the evaluating group against the evaluated adapted model, or confined the cell to the developers' own evaluation, where the first rater had recorded independence from the backbone developers or the collaborating sites (D+C against D, 5 cells; I against D, 3; D+I against D, 2; D+C against I, 1). The raters marked preprints by different conventions (the first only for identified cells), so agreement on the preprint mark is informative only for identified cells, where it was 23 of 25; the conventional-context mark agreed in 105 of 116 cells.

Adjudication. The 15 state disagreements (Table S17) were referred to a third author (S.Z.M.A.) on a form that withheld both raters' codes. The entries in his returned form, 14 U and one N, are identical in wording to adjudication proposals drafted with AI assistance from records that included both raters' codes and notes, which he reviewed and approved. The third review was therefore not blinded and was not independent of the drafted proposals; we describe it as review and approval of adjudication proposals, not as an independent third rating. For two cells all three codes differed (BB1-B1 and U3-A2: N, V and U); both were resolved as U. S.Z.M.A. confirmed that he reviewed and approved the adjudication proposals for all 15 disputed evidence states. All authors confirmed their approval of the final 116-cell consensus matrix and the 11 disputed evaluator-code decisions, including the six changes to existing evaluator codes. The consensus was approved on 6 November 2026. The original rater classifications were preserved separately, and the third review remains characterized as approval of AI-assisted proposals rather than a blinded independent third assessment.

Effect of adjudication. Consensus changed 11 states: four from V to U (U4-A1, U5-A1, U1-A4, U1-A7) and seven from N to U (BB1-B1, BB1-B6, BB2-B6, U1-A2, U3-A2, U2-A3, U2-A12); four disputed cells kept their original state (U4-A7, U5-A7 and U4-A9 as U; U1-A10 as N). The V to U changes apply the rule of Table S6 strictly (values must agree across sections and versions, and required design elements must be stated in the report); the N to U changes record candidate evidence found by the second rater that could not be verified. The two B6 cells rest on a preprint whose full text none of the raters or the adjudicator could read, so they are holding codes, not findings that a qualifying experiment exists. Six evaluator codes changed: D+C to D for BB3-B2, BB3-B3 and U4-A8, and I to D (backbone-independent) for U2-A2, U3-A1 and U3-A3. Evaluator codes for the seven cells that changed from N to U were assigned by applying the same rule and were not decided separately. Under this rule no identified adapted-model cell rests on evaluators independent of the adapted model's developers;

the original text reported three of nineteen, counting independence from the backbone developers.

What the analysis shows. Two authors applying the coding guide to the same sources agreed on most states, and their disagreements lay at the V/U and U/N boundaries and in the evaluator codes, which the original text had identified as the most judgement-dependent elements; every whole row coded N (intervention, independent locked validation, test–retest) was coded N by both raters. That agreement, including 35 of 35 intervention cells, shows that neither rater found such evidence in the sources examined; it does not show that none exists or test how intervention evidence would be coded. The analysis does not show that the codes are correct, that raters outside the author team would agree, or that the framework or its principles are valid. Agreement may be overstated because both raters are authors who share the framework's assumptions and both worked with AI-assisted extraction.

Table S16. Agreement between the two raters on the evidence state before adjudication (original codes; V, U, N, X).

Analysis set	Cells	Agreement, n (%)	Cohen's κ	Gwet's AC1	Comment
All cells (principal analysis)	116	101 (87.1)	0.759	0.843	Four categories
Planned analysis set (excludes the 26 planned training cells)	90	77 (85.6)	0.728	0.825	Sensitivity analysis; no training took place
Planned training cells (ECG-FM backbone; structural heart disease use)	26	24 (92.3)	0.861	0.906	Coded independently like all other cells
Backbone level	21	18 (85.7)	0.777	0.819	
Adapted-model level	60	48 (80.0)	0.656	0.752	
Intervention level	35	35 (100)	Undefined	1.000	Both raters used only N; AC1 trivial
Excluding cells both raters coded X	110	95 (86.4)	0.720	0.820	Three categories (V, U, N)
Identified (V or U) versus N versus X	116	108 (93.1)	0.861	0.908	V and U merged
Evaluator code, cells with the same identified state	25	14 (56.0)	Not computed	Not computed	Few cells per category

Rater 1, M.A.M. (original codes); rater 2, R.Q. Cohen's κ uses each rater's marginal distribution; Gwet's AC1 uses chance agreement $\Sigma\pi k(1-\pi k)/(q-1)$ with π the mean of the two raters' proportions and q the number of categories. No confidence intervals are given because cells are clustered within 8 objects and 26 requirements (Supplementary Note 5). Confusion matrices and calculations are in Supplementary Data 2.

Table S17. Adjudication of the 15 state disagreements.

Cell	Requirement	Rater 1	Rater 2	Third review	Consensus	Main reason
BB1-B1	B1	N	V	U	U	Evaluation sets labelled external, but no exclusion statement or overlap audit in the text read; possible CODE membership of the pretraining corpus unresolved
BB1-B6	B6	N	U	U	U	Candidate robustness preprint (SSRN 6753359) identified by rater 2; full text not accessible; holding code
BB2-B6	B6	N	U	U	U	As BB1-B6
U4-A1	A1	V	U	U	U	Matched supervised comparator present, but reported values differ between sections and versions

Cell	Require-ment	Rater 1	Rater 2	Third re-view	Consensus	Main reason
U5-A1	A1	V	U	U	U	Values differ between sections and versions; comparator initialized from a supervised model
U1-A2	A2	N	U	U	U	Pretraining corpus named, but no statement that the evaluation hospitals lie outside it
U3-A2	A2	N	V	U	U	Separation inferable only by combining two reports; the evaluating report does not state it
U2-A3	A3	N	U	U	U	Cross-validation grouped by recording; patient-level separation not stated
U1-A4	A4	V	U	U	U	Model fixed, but transport of the threshold to the external site not stated
U1-A7	A7	V	U	U	U	External operating point at the primary threshold incomplete (NPV only); threshold transport not stated
U4-A7	A7	U	V	U	U	Supplementary tables do not state which threshold produced the reported values
U5-A7	A7	U	V	U	U	Thresholds listed for the internal test set only; which threshold produced external values not stated
U4-A9	A9	U	V	U	U	Pooled disparity bounds only; group-specific performance and thresholds not verified
U1-A10	A10	N	U	N	N	Platforms reported at set level only; no stratified or degradation analysis
U2-A12	A12	N	U	U	U	Per-form performance with a separately fitted classifier for each form; equivalence of one model not shown

Rater 1, M.A.M.; rater 2, R.Q.; third review, S.Z.M.A. (review and approval of AI-drafted adjudication proposals prepared from records that included both raters' codes; not blinded). Consensus approved by all authors on 6 November 2026. Sources and locations for each decision are in Table S11 and Supplementary Data 2.

References

1. Ye, J., S. Bronstein and M. A. Hashish. "Foundation models for cardiovascular diseases and medicine." *European Heart Journal - Digital Health* 7 (2026): ztag113. 10.1093/ehjdh/ztag113. https://doi.org/10.1093/ehjdh/ztag113.
2. Chen, H., J. Huang and X. Hou. "Translational maturity of artificial intelligence-enabled electrocardiography varies across cardiovascular diseases." *Frontiers in Cardiovascular Medicine* 13 (2026): 1940621. 10.3389/fcvm.2026.1940621. https://doi.org/10.3389/fcvm.2026.1940621.
3. Han, Y., V. Murino, X. Liu, X. Zhang and C. Ding. "A systematic review on foundation models for electrocardiogram analysis: Initial strides and expansive horizons." *arXiv* (2024): 10.48550/arXiv.2410.19877. https://arxiv.org/abs/2410.19877.
4. Khan, M. H., A. Nayfeh, M. Masood, A. A. Al-Shaikhi, M. M. U. Rahman and T. Y. Al-Naffouri. "Ecg foundation models and medical llms for agentic cardiovascular intelligence at the edge: A review and outlook." *arXiv* (2026): 10.48550/arXiv.2604.02501. https://arxiv.org/abs/2604.02501.
5. Al-Masud, M. A., J. M. L. Alcaraz and N. Strodthoff. "Benchmarking ecg fms: A reality check across clinical tasks." International Conference on Learning Representations (ICLR 2026), 2026.

6. Nolin-Lapalme, A., A. Sowa, J. Delfrate, O. Tastet, D. Corbin, M. Kulbay, D. Ozdemir, M.-J. Noël, F.-C. Marois-Blanchet, F. Harvey, *et al.* "Foundation models for electrocardiogram interpretation: Clinical implications." *European Heart Journal* 47 (2026): 2174-86. 10.1093/eurheartj/ehaf1119. https://doi.org/10.1093/eurheartj/ehaf1119.
7. Nolin-Lapalme, A., A. Sowa, J. Delfrate, O. Tastet, D. Corbin, M. Kulbay, D. Ozdemir, M.-J. Noel, F.-C. Marois-Blanchet, F. Harvey, *et al.* "Foundation models for generalizable electrocardiogram interpretation: Comparison of supervised and self-supervised electrocardiogram foundation models." *medRxiv* (2025): 10.1101/2025.03.02.25322575. https://doi.org/10.1101/2025.03.02.25322575.
8. Yang, Y., L. Peracchio, J. Mayourian, T. Miller and W. La Cava. "An ecg foundation model for generalizable cardiac function prediction across the lifespan." *medRxiv* (2026): 10.64898/2026.05.26.26354128. https://doi.org/10.64898/2026.05.26.26354128.
9. Kim, S. "Learning general representation of 12-lead electrocardiogram with a joint-embedding predictive architecture." *arXiv* (2024): 10.48550/arXiv.2410.08559. https://arxiv.org/abs/2410.08559.
10. Coppola, E., M. Savardi, M. Massussi, M. Adamo, M. Metra and A. Signoroni. "Hubert-ecg as a self-supervised foundation model for broad and scalable cardiac applications." *medRxiv* (2024): 10.1101/2024.11.14.24317328. https://doi.org/10.1101/2024.11.14.24317328.
11. Filice, F., E. De Rose, S. Bartucci, F. Calimeri and S. Perri. "Looking beyond accuracy: A holistic benchmark of ecg foundation models." *arXiv* (2026): 10.48550/arXiv.2601.21830. https://arxiv.org/abs/2601.21830.
12. Scagnetto, A. "Ecg biometrics with arcface-inception: External validation on mimic and heedb." *arXiv* (2026): 10.48550/arXiv.2604.04485. https://arxiv.org/abs/2604.04485.
13. Li, J., A. D. Aguirre, V. Moura, J. Jin, C. Liu, L. Zhong, C. Sun, G. D. Clifford, M. B. Westover and S. Hong. "An electrocardiogram foundation model built on over 10 million recordings." *NEJM AI* 2 (2025): 10.1056/AIoa2401033. https://doi.org/10.1056/AIoa2401033.
14. Lee, S., J.-W. Son, S.-A. Kim, M.-S. Ahn, S. J. Lee, S.-J. Han, T. Joo, Y. Na, S. Joo, H. J. Ahn, *et al.* "Deep learning model using transfer learning for detecting left ventricular systolic dysfunction: Retrospective algorithm development and validation study." *JMIR Medical Informatics* 14 (2026): e83127. 10.2196/83127. https://doi.org/10.2196/83127.
15. Tang, G., Q. Zhao, G. Nie, Y. Xiao, S. Geng, D. Xie, S. Huang, D. Zhang, X. Yao, J. Wang, *et al.* "Artificial intelligence-enabled single-lead ecg for non-invasive hyperkalemia detection: Development, multicenter validation, and proof-of-concept deployment." *arXiv* (2026): 10.48550/arXiv.2603.14177. https://arxiv.org/abs/2603.14177.
16. McKeen, K., S. Masood, A. Toma, B. Rubin and B. Wang. "Ecg-fm: An open electrocardiogram foundation model." *JAMIA Open* 8 (2025): ooaf122. 10.1093/jamiaopen/ooaf122. https://doi.org/10.1093/jamiaopen/ooaf122.
17. Monachino, G., B. Zanchi, G. Farina and F. D. Faraci. "Beyond full fine-tuning: Towards enhanced generalizability in downstream ecg foundation model adaptation." *Scientific Reports* (2026): 71351. 10.1038/s41598-026-71351-2. https://doi.org/10.1038/s41598-026-71351-2.
18. Monachino, G., N. La Porta, B. Zanchi, L. Fiorillo, A. D. Rossi, G. Farina and F. D. Faraci. "Self-dana: A resource-efficient channel-adaptive self-supervised approach for ecg foundation models." *arXiv* (2025): 10.48550/arXiv.2507.14151. https://arxiv.org/abs/2507.14151.
19. Dhingra, L. S., P. M. Croon, B. Batinica, E. K. Oikonomou and R. Khera. "Test-retest reliability of artificial intelligence-enhanced electrocardiography: A multi-center study." *medRxiv* (2025): 10.1101/2025.11.04.25339526. https://doi.org/10.1101/2025.11.04.25339526.
20. Hsieh, P. N., P. Agrawal, A. Alok, S. Kumar, C. Ramanathan, V. L. Murthy, N. Varma, V. Nagarajan, A. P. Ambrosy, M. Ramsis, *et al.* "Artificial intelligence electrocardiography for left ventricular systolic dysfunction

demonstrates preserved performance across demographic training imbalances." *European Heart Journal - Digital Health* 7 (2026): ztag080. 10.1093/ehjdh/ztag080. https://doi.org/10.1093/ehjdh/ztag080.

21. Noseworthy, P. A., Z. I. Attia, L. C. Brewer, S. N. Hayes, X. Yao, S. Kapa, P. A. Friedman and F. López-Jiménez. "Assessing and mitigating bias in medical artificial intelligence: The effects of race and ethnicity on a deep learning model for ecg analysis." *Circulation: Arrhythmia and Electrophysiology* 13 (2020): e007988. 10.1161/CIRCEP.119.007988. https://doi.org/10.1161/CIRCEP.119.007988.
22. Nguyen, H. D., T.-T. Pham, N. Le and V. Nguyen. "Tolerantecg: A foundation model for imperfect electrocardiogram." *arXiv* (2025): 10.48550/arXiv.2507.09887. https://arxiv.org/abs/2507.09887.
23. Taleshinosrati, A., Y. Wang, A. Phoemsuk, V. Abolghasemi, N. H. Motlagh, S. Puthusserypady, D. Teichmann and A. Peimankar. "Found-af: Benchmarking ecg foundation models for atrial fibrillation detection." *arXiv* (2026): 10.48550/arXiv.2608.03597. https://arxiv.org/abs/2608.03597.
24. Do, D. N., M. N. Do, D. H. Nguyen, K. T. Q. Le, K. D. Pham, H. N. Huynh, P. Pham-Van-Hoang, Q. K. Huynh, R. M. Odat, P. Ashar, *et al.* "Domain-adapted fine-tuning of ecg foundation models for multi-label structural heart disease screening." *arXiv* (2026): 10.48550/arXiv.2604.23385. https://arxiv.org/abs/2604.23385.
25. Lunelli, R., A. Nicolson, S. M. Pröll, S. J. Reinstadler, A. Bauer and C. Dlaska. "BenchECG and xECG: A benchmark and baseline for ECG foundation models." *npj Digital Medicine* (2026): 10.1038/s41746-026-03196-y. https://doi.org/10.1038/s41746-026-03196-y.
26. Al-Masud, M. A. and N. Strodthoff. "Pretraining strategies and scaling for ecg foundation models: A systematic study." *arXiv* (2026): 10.48550/arXiv.2605.12241. https://arxiv.org/abs/2605.12241.
27. Bollmann, A., I. Kim, S. Hohenstein, M. Nitsche, L. Koehler, M. Schemmer, P. Schmitz, V. Pernstecher, S. Koenig, J. Leiner, *et al.* "Transformer-based ecg foundation and deep learning models for assessment of left ventricular systolic dysfunction." *European Heart Journal - Digital Health* 7 (2026): 10.1093/ehjdh/ztaf143.045. https://doi.org/10.1093/ehjdh/ztaf143.045.
28. Na, Y., M. Park, Y. Tae and S. Joo. "Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram." International Conference on Learning Representations (ICLR 2024), 2024. 15012-35.
29. Liu, C., Z. Wan, C. Ouyang, A. Shah, W. Bai and R. Arcucci. "Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement." Proceedings of the 41st International Conference on Machine Learning, PMLR, 2024. 235, 31949-63.
30. Wang, F., J. Xu and L. Yu. "From token to rhythm: A multi-scale approach for ecg-language pretraining." International Conference on Machine Learning (ICML 2025), 2025.
31. Ko, M., M. Gadaleta, E. J. Topol, E. D. Muse and G. Quer. "Development and external validation of a contrastive learning foundation model for ecg-based prediction of cardiovascular diseases and outcomes." *The Lancet Digital Health* (2026): 101092. 10.1016/j.landig.2026.101092. https://doi.org/10.1016/j.landig.2026.101092.
32. Hu, S. M., J. P. Barrios and G. H. Tison. "A deep foundation model for electrocardiogram interpretation: Enabling rare disease detection through transfer learning." *European Heart Journal - Digital Health* 6 (2025): 619-23. 10.1093/ehjdh/ztaf051. https://doi.org/10.1093/ehjdh/ztaf051.
33. Diamant, N., E. Reinertsen, S. Song, A. D. Aguirre, C. M. Stultz and P. Batra. "Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling." *PLoS Computational Biology* 18 (2022): e1009862. 10.1371/journal.pcbi.1009862. https://doi.org/10.1371/journal.pcbi.1009862.
34. Xiao, Y., Q. Zhao, G. Tang, H. Zhang, Z. Kan, D. Zhang, J. Li, G. Nie, X. Fang, H. Wang, *et al.* "Fine-tuning an ecg foundation model to predict coronary CT angiography outcomes." *npj Digital Medicine* (2026): 10.1038/s41746-026-03085-4. https://doi.org/10.1038/s41746-026-03085-4.

35. Carter, R. E., P. W. Johnson, J. B. Strom, J. W. Waks, A. Krumerman, K. J. Ferrick, R. DeRaad, B. A. Steinberg, M. A. Wieczorek, J. Cruz, *et al.* "Multisite, external validation of an AI-enabled ecg algorithm for detection of low ejection fraction." *JACC: Advances* 5 (2026): 102537. 10.1016/j.jacadv.2025.102537. https://doi.org/10.1016/j.jacadv.2025.102537.

36. Lee, H., Y. Kim, D. Sykora, A. J. Ryu, Y. Cho, J. Kim and J. Song. "External validation of ecg artificial intelligence for emergency and cardiac assessment across a large-scale u.S. Healthcare system." *npj Digital Medicine* 9 (2026): 509. 10.1038/s41746-026-02682-7. https://doi.org/10.1038/s41746-026-02682-7.

37. Sau, A., L. Pastika, E. Sieliwonczyk, K. Patlatzoglou, A. H. Ribeiro, K. A. McGurk, B. Zeidaabadi, H. Zhang, K. Macierzanka, D. P. Mandic, *et al.* "Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: A model development and validation study." *The Lancet Digital Health* 6 (2024): e791-e802. 10.1016/S2589-7500(24)00172-9. https://doi.org/10.1016/S2589-7500(24)00172-9.

38. Sangha, V., B. J. Mortazavi, A. D. Haimovich, A. H. Ribeiro, C. Brandt, D. L. Jacoby, W. L. Schulz, H. M. Krumholz, A. L. P. Ribeiro and R. Khera. "Automated multilabel diagnosis on electrocardiographic images and signals." *Nature Communications* 13 (2022): 1583. 10.1038/s41467-022-29153-3. https://doi.org/10.1038/s41467-022-29153-3.

39. Yao, X., D. Rushlow, J. Inselman, R. G. McCoy, T. D. Thacher, E. Behnken, M. Bernard, S. Rosas, A. Akfaly, A. Misra, *et al.* "Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: A pragmatic, randomized clinical trial." *Nature Medicine* 27 (2021): 815-19. 10.1038/s41591-021-01335-4. https://doi.org/10.1038/s41591-021-01335-4.

40. Rushlow, D. R., I. T. Croghan, J. W. Inselman, T. D. Thacher, P. A. Friedman, X. Yao, P. A. Pellikka, F. Lopez-Jimenez, M. E. Bernard, B. A. Barry, *et al.* "Clinician adoption of an artificial intelligence algorithm to detect left ventricular systolic dysfunction in primary care." *Mayo Clinic Proceedings* 97 (2022): 2076-85. 10.1016/j.mayocp.2022.04.008. https://doi.org/10.1016/j.mayocp.2022.04.008.

41. Thao, V., Y. Zhu, A. S. Tseng, J. W. Inselman, B. J. Borah, R. G. McCoy, Z. I. Attia, F. Lopez-Jimenez, P. A. Pellikka, D. R. Rushlow, *et al.* "Cost-effectiveness of artificial intelligence-enabled electrocardiograms for early detection of low ejection fraction: A secondary analysis of the electrocardiogram artificial intelligence-guided screening for low ejection fraction trial." *Mayo Clinic Proceedings: Digital Health* 2 (2024): 620-31. 10.1016/j.mcpdig.2024.10.001. https://doi.org/10.1016/j.mcpdig.2024.10.001.

42. Tsai, D.-J., C. Lin, W.-T. Liu, C.-C. Lee, C.-H. Chang, W.-Y. Lin, Y.-L. Liu, D.-W. Chang, P.-H. Hsieh, C.-S. Tsai, *et al.* "Artificial intelligence-assisted diagnosis and prognostication in low ejection fraction using electrocardiograms in inpatient department: A pragmatic randomized controlled trial." *BMC Medicine* 23 (2025): 342. 10.1186/s12916-025-04190-z. https://doi.org/10.1186/s12916-025-04190-z.

43. Lin, C.-S., W.-T. Liu, D.-J. Tsai, Y.-S. Lou, C.-H. Chang, C.-C. Lee, W.-H. Fang, C.-C. Wang, Y.-Y. Chen, W.-S. Lin, *et al.* "AI-enabled electrocardiography alert intervention and all-cause mortality: A pragmatic randomized clinical trial." *Nature Medicine* 30 (2024): 1461-70. 10.1038/s41591-024-02961-4. https://doi.org/10.1038/s41591-024-02961-4.

44. Lin, C., W.-T. Liu, C.-H. Chang, C.-C. Lee, S.-C. Hsing, W.-H. Fang, D.-J. Tsai, K.-C. Chen, C.-H. Lee, C.-C. Cheng, *et al.* "Artificial intelligence-powered rapid identification of st-elevation myocardial infarction via electrocardiogram (arise): A pragmatic randomized controlled trial." *NEJM AI* 1 (2024): 10.1056/AIoa2400190. https://doi.org/10.1056/AIoa2400190.

45. Lin, C., C.-S. Lin, S.-J. Chen, S.-H. Tsai, C.-C. Sung, C.-C. Chen, Y.-J. Hsu, Y.-J. Hung and S.-H. Lin. "AI-enabled electrocardiogram alert for potassium imbalance treatment: A pragmatic randomized controlled trial." *Nature Communications* 17 (2026): 159. 10.1038/s41467-025-66394-4. https://doi.org/10.1038/s41467-025-66394-4.

46. Adedinsewo, D., A. C. Morales-Lara, B. B. Afolabi, O. A. Kushimo, A. Mbakwem, K. F. Ibiyemi, J. A. Ogunmodede, H. O. Raji, S. H. Ringim, A. A. Habib, *et al.* "Artificial intelligence guided screening for

cardiomyopathies in an obstetric population: A pragmatic randomized clinical trial." *Nature Medicine* 30 (2024): 2897-906. 10.1038/s41591-024-03243-9. https://doi.org/10.1038/s41591-024-03243-9.

47. Kelshiker, M. A., P. Bächtiger, C. F. Petri, S. Nakhare, J. Mansell, K. Chhatwal, A. Alrumayh, J. Zaman, M. Shah, H. Young, *et al.* "Triple cardiovascular disease detection with an artificial intelligence-enabled stethoscope (tricorder) in the uk: A cluster-randomised controlled implementation trial." *The Lancet* 407 (2026): 704-15. 10.1016/S0140-6736(25)02156-7. https://doi.org/10.1016/S0140-6736(25)02156-7.
48. Simonetto, D. A., D. Rushlow, K. Liu, A. Calleri, B. A. Kassmeyer, R. J. Lennon, P. Rattan, M. E. Bernard, G. Singh, M. E. Deyo-Svendsen, *et al.* "Detection of undiagnosed liver cirrhosis via AI-enabled electrocardiogram: A pragmatic, cluster-randomized clinical trial." *Nature Medicine* 32 (2026): 160-67. 10.1038/s41591-025-04058-y. https://doi.org/10.1038/s41591-025-04058-y.
49. Noseworthy, P. A., Z. I. Attia, E. Behnken, R. Giblon, K. A. Bews, S. Liu, T. Gosse, Z. D. Linn, Y. Deng, J. Yin, *et al.* "Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: A prospective non-randomised interventional trial." *The Lancet* 400 (2022): 1206-12. 10.1016/S0140-6736(22)01637-3. https://doi.org/10.1016/S0140-6736(22)01637-3.
50. Calleri, A., Y. Yazarkan, K. Liu, B. A. Kassmeyer, R. J. Lennon, P. Rattan, A. Seid, M. E. Bernard, G. Singh, M. E. Deyo-Svendsen, *et al.* "Clinician engagement shapes the impact of AI-based ecg screening for chronic liver disease in primary care." *npj Digital Medicine* 9 (2026): 10.1038/s41746-026-02718-y. https://doi.org/10.1038/s41746-026-02718-y.
51. Vedage, N. A., S. F. Friedman, Y. Chang, L. H. Borowsky, S. J. Shah, D. D. McManus, S. J. Atlas, D. E. Singer, S. A. Lubitz, M. Maddah, *et al.* "Risk-guided atrial fibrillation screening with artificial intelligence-enabled electrocardiogram models: A vital-af trial analysis." *Journal of the American College of Cardiology* 87 (2026): 1798-813. 10.1016/j.jacc.2026.01.087. https://doi.org/10.1016/j.jacc.2026.01.087.
52. Khurshid, S., S. F. Friedman, M. A. Al-Alusi, S. Kany, T. F. Sommers, C. D. Anderson, J. E. Ho, D. D. McManus, L. H. Borowsky, J. M. Ashburner, *et al.* "Artificial intelligence-enabled analysis of handheld single-lead electrocardiograms to predict incident atrial fibrillation: An analysis of the vital-af randomized trial." *npj Digital Medicine* 8 (2025): 776. 10.1038/s41746-025-02164-2. https://doi.org/10.1038/s41746-025-02164-2.
53. Ong, A. Q. C., C.-S. Ang, I. Bojić, C. L. Johnson, H. Aggour, F. S. Ng, J. Car, P. Leeson, J. Gonçalves and N. M. Lai. "Artificial intelligence in cardiovascular care: A systematic review and meta-analysis of randomised controlled trials." *eClinicalMedicine* 98 (2026): 104123. 10.1016/j.eclinm.2026.104123. https://doi.org/10.1016/j.eclinm.2026.104123.
54. Hadida Barzilai, D., K. Sudri, G. Goshen, E. Klang, E. Zimlichman, I. M. Barbash and M. Cohen Shelly. "Randomized controlled trials evaluating artificial intelligence in cardiovascular care: A systematic review." *JACC: Advances* 4 (2025): 102152. 10.1016/j.jacadv.2025.102152. https://doi.org/10.1016/j.jacadv.2025.102152.
55. Hospital, T. "Effect of smartwatch along with foundation model for the monitoring and management of acute myocardial infarction patients during outpatient cardiac rehabilitation after percutaneous coronary intervention: Protocol for a randomised controlled multicentre trial (nct07395648)." *ClinicalTrials.gov* (2026): https://clinicaltrials.gov/study/NCT07395648.
56. Institute, M. H. "Harnessing ecg artificial intelligence for rapid treatment and accurate interpretation, an open label randomized controlled trial (heart-AI; nct06462989)." *ClinicalTrials.gov* (2024): https://clinicaltrials.gov/study/NCT06462989.
57. Institute, M. H. "Deployment and evaluation of artificial intelligence software for electrocardiogram analysis and management in primary care (daisea-ecg; nct06637293)." *ClinicalTrials.gov* (2024): https://clinicaltrials.gov/study/NCT06637293.

58. Ambulanciers, É. S. d. S. "AI & prehospital ecg analysis: A randomized controlled trial of a smartphone large language model for occlusion myocardial infarction detection by prehospital providers (ecg-ia; nct07810686)." *ClinicalTrials.gov* (2026): https://clinicaltrials.gov/study/NCT07810686.

59. Xu, Y., J. Lu, S. Ding, D. Cao, X. Hu and C. Yang. "An electrocardiogram multi-task benchmark with comprehensive evaluations and insightful findings." *arXiv* (2025): 10.48550/arXiv.2512.08954. https://arxiv.org/abs/2512.08954.

60. Muneer, A.; Waqas, M.; Saad, M.B.; Showkatian, E.; Bandyopadhyay, R.; Xu, H.; Li, W.; Chang, J.Y.; Liao, Z.; Haymaker, C.; et al. From classical machine learning to emerging foundation models: review on multimodal data integration for cancer research. *Artif. Intell. Rev.* 2026, 59, 119.

61. Duan, Y. and W. Qiu. "ECG-InterpBench: Benchmarking the interpretability of ECG foundation models with matched-scale sparse autoencoders." *arXiv* (2026): 10.48550/arXiv.2607.27404. https://arxiv.org/abs/2607.27404.

62. Abrar, M. F., M. Saqib, S. Ali, I. E. Lee, A. Khan, S. Jan and A. Alkhayyat. "Automated heart disease detection using swin transformer and ecg signal processing: A high-accuracy approach." *Scientific Reports* 15 (2025): 39338. 10.1038/s41598-025-23097-6. https://doi.org/10.1038/s41598-025-23097-6.