TARGET-AI: a foundational approach for the targeted deployment of artificial intelligence electrocardiography in the electronic health record

Evangelos K. Oikonomou, MD, DPhil[a,b*], Bruno Batinica, MBBS[a,b], Lovedeep S. Dhingra, MBBS[a,b], Arya Aminorroaya, MD, MPH[a,b], Andreas Coppi, PhD,[b,c] Rohan Khera, MD, MS[a,b,c,d,e*]

a. Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA.
b. Cardiovascular Data Science (CarDS) Lab, Yale School of Medicine, New Haven, CT, USA.
c. Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, USA.
d. Section of Biomedical Informatics and Data Science, Yale School of Medicine, New Haven, CT, USA.
e. Section of Health Informatics, Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA.

Manuscript type: Original research
Total word count (main text): 3,165 words
Main figures: 5; **main tables:** 1
Funding: The authors acknowledge support by the National Heart, Lung, And Blood Institute of the National Institutes of Health (under award numbers R01HL167858 and K23HL153775 to RK, and F32HL170592 to EKO), the National Institute on Aging of the National Institutes of Health (under award number R01AG089981 to RK), and the Doris Duke Charitable Foundation (under award number 2022060 to RK). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

***Correspondence:**
Evangelos K. Oikonomou, MD, DPhil
195 Church St, 6th Floor, New Haven, CT 06510
evangelos.oikonomou@yale.edu; @ekoikonomou

Rohan Khera, MD, MS
195 Church St, 6th Floor, New Haven, CT 06510
rohan.khera@yale.edu; @rohan_khera

ABSTRACT

Background: Artificial intelligence (AI) applied to routine electrocardiograms (ECGs) offers promise for screening of structural heart disease (SHD), yet broad clinical integration remains limited by high false positive rates and the lack of tailored deployment strategies.

Methods: We developed TARGET-AI, a multimodal AI-enabled pipeline that integrates longitudinal electronic health record (EHR) data with ECG images to identify optimal intersections of healthcare encounters and patient phenotypes for targeted AI-ECG screening. The approach is built on (1) a foundation model pretrained on 118 million coded EHR events from 159,322 individuals to generate temporal patient embeddings and identify high-risk screening candidates, followed by (2) a contrastive vision-language model trained on 754,533 ECG-echocardiogram pairs to detect SHD with tunable performance characteristics. We evaluated this joint strategy in a temporally distinct cohort of 5,198 individuals referred for their first transthoracic echocardiogram (TTE) within 90 days of an ECG and externally in 33,518 participants from the UK Biobank undergoing ECG and cardiac magnetic resonance imaging.

Results: Our pre-trained AI-ECG image foundation model discriminated 27 SHD subtypes, from left ventricular systolic dysfunction (AUROC of 0.90) to severe aortic stenosis (AUROC of 0.85) and elevated right ventricular systolic pressure (AUROC of 0.82). Compared with untargeted AI-ECG screening, EHR-informed TARGET-AI-guided screening significantly reduced false positive predictions across SHD labels (median reduction: 87.8%; interquartile range [IQR], 82.4%-98.2%) and improved F1 score (median increase: 0.25; IQR, 0.19-0.41). In the UK Biobank, targeted screening reduced false positives by 61.7% (IQR, 50.4%-89.1%) while preserving sensitivity.

Conclusions: TARGET-AI enables the context-aware deployment of AI-ECG screening by leveraging key longitudinal EHR phenotypes and multimodal ECG-echocardiogram representations, thereby defining an interoperable, data-driven strategy for the more precise deployment of AI screening tools across health systems.

KEYWORDS: electronic health record, artificial intelligence, electrocardiography, screening, implementation

INTRODUCTION

Artificial intelligence (AI) technologies hold promise for the scalable detection of structural heart disease (SHD) from routine cardiac diagnostic tests, such as traditional 12-lead electrocardiograms (ECG).[1,2] Nevertheless, evidence remains limited on how best to integrate these tools into clinical care pathways.[3,4] Two key restrictions, in particular, limit the uptake of these technologies within the electronic health records (EHR) of large health systems. First, the prevalence of SHD in unselected cohorts is low, leading to high false discovery rates,[5] unnecessary testing, patient anxiety, and increased downstream costs.[6–8] Second, any system intended to guide the targeted deployment of AI-ECG technologies for cardiac screening must exhibit foundational capabilities that allow it to be flexibly adapted to different forms of SHD and diverse patient populations.[9] Despite the emergence of foundation models that learn global representations of health and disease across modalities,[10] including ECG,[11] transthoracic echocardiography (TTE),[12] and EHR records,[13,14] there is currently no systematic strategy that integrates these data sources to dynamically estimate the pre-test probability of SHD and guide the targeted deployment of AI-ECG.

To bridge this gap, we introduce *TARGET-AI*, a foundational, multi-modal approach that learns interoperable, longitudinal representations from the EHR, alongside ECG image embeddings that have been contrastively pre-trained against linked echocardiograms, to guide the optimal deployment of AI-ECG in health systems. We first developed this system in a cohort of more than 150,000 unique individuals, contributing more than 18 million EHR events and more than 700,000 ECG-TTE pairs. We then tested it as a strategy for optimal deployment of AI-ECG in an independent, temporally distinct set of individuals as well as externally in the UK Biobank. Consistent with the overarching goal of reducing false discovery, we evaluated the system's ability to improve the balance of precision and recall across multiple SHD labels that are not routinely detectable by ECG alone. Conceptually, this enabled us to evaluate two alternative clinical deployment strategies for AI-ECG (**Fig. 1**): (i) an *untargeted* approach, in which models are applied opportunistically to all individuals with available ECG data; and (ii) a *targeted* strategy, in which deployment is guided by data-driven estimates of disease likelihood derived from each individual's longitudinal EHR trajectory and spot ECG features.

METHODS

Data Source and Population

This retrospective study was approved by the Yale Institutional Review Board (IRB) with a waiver of informed consent. UK Biobank analyses were conducted under project #71033.

An extended version of the methods can be found in the **Supplement**. Briefly, we queried the Yale-New Haven Health System (YNHHS) EHR (2013-2023) to identify adults (≥18 years) who underwent a 12-lead ECG and TTE within 90 days (±). In total, we assembled 754,533 ECG-TTE pairs from 159,322 individuals (*2013-2021*) for model training, with a cap set at 10 ECGs per person. A held-out test set included 8,979 ECG-TTE pairs from distinct individuals (*2022-2023*). To model a temporally valid screening strategy, we further restricted this subset to 5,198 patients undergoing a first TTE within 90 days after the index ECG.

In the UK Biobank, we identified 33,518 participants with protocolized 12-lead ECG and cardiac magnetic resonance (CMR) imaging (assessment visit #2). Demographic and clinical data were extracted from structured fields and protocol-derived labs. Imaging labels were based on automated CMR measurements, using established thresholds for key parameters, such as LVEF ≤40%, or cohort-specific 99th percentiles for volumetric and linear measurements.[15] The detailed step-to-step analytical framework, definitions and approach are summarized in the ***Supplement***.

Data Preprocessing and Label Definition

In all cohorts ECG images were generated at random across four layouts (standard, alternate, standard shuffled and alternate shuffled) to enhance generalizability, as described and validated in our previous work.[16–19] In YNHHS, unstructured (textual) and structured TTE reports were exported from the local reporting software (Lumedx, Oakland, CA).[20] To cover the broad spectrum of SHD phenotypes, we defined 27 discrete SHD labels, spanning parameters of left ventricular and right ventricular function, size, thickness, atrial anatomy, valvular function and related measurements (**Table S1**). In the UK Biobank, CMR-derived phenotypes were computed using automated fields from Bai et al.[15] SHD definitions incorporated standard cutoffs and population-based thresholds (see ***Supplement***).

CLMBR-T Embeddings

To encode longitudinal EHR representations we used CLMBR-T (**Fig. S1**), a 141M parameter autoregressive foundation model pre-trained on 2.57 million EHR records from Stanford Medicine.[13,14] CLMBR-T models a sequence of coded medical events, spanning diagnoses, medications, laboratory measurements, procedures, and encounter history, based on the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) vocabulary and MEDS (Medical Event Data Standard) schema. For each ECG, a MEDS-compatible event sequence was generated using all events recorded in the EHR from its inception until the timepoint the ECG was performed. We extracted the final 768-dimensional hidden layer as a representation of a patient's longitudinal profile. In total, we successfully mapped 26,828 [96.0%] of the original 27,946 non-laboratory-related codes using Unified Medical Language System (UMLS) resources (e.g., SNOMED CT, RxNorm) (**Tables S2-S3**). In addition to age, sex, race and ethnicity, these spanned a comprehensive list of 1,982 diagnosis concepts, 213 medications, 25,771 medical procedures, and 37 commonly performed laboratory measurements across 3 possible encounter settings (inpatient, ED/observation, outpatient). More details on the definitions and mapping schemas used across the YNHHS and UK Biobank cohorts are provided in the ***Supplement***.

Contrastive ECG-TTE Vision-Text Model Training

To learn echocardiography-defined SHD representations from ECG images, we trained a CLIP (Contrastive Language-Image Pretraining)-style model to align ECG images with TTE reports using a dual encoder architecture (**Fig. 2a**). First, we defined patient-level splits (95% train, 5% validation), with up to 10 ECGs per patient and randomized layout augmentation. Next, we initialized our ECG image (vision) encoder using BEiT-base-patch16-384 ViT (vision transformer; 12 layers, 768 hidden dimensions),[21] whereas for the text encoder we used a CLIP text transformer (512 hidden dimensions, 256

tokens). To account for the limited vocabulary found in TTE reports, we trained a custom BPE (byte-pair encoding) tokenizer with a final vocabulary size of 16,030 and a minimum token frequency of 3. We trained the contrastive model in up to 8 NVIDIA H100 graphics processing units (GPUs) for a total of 100 epochs using the AdamW optimizer, a 10^{-5} learning rate, linear warm-up and a cosine decay schedule. We also implemented mixed-precision training to reduce memory usage, using a batch size of 64 and gradient accumulation steps set to 1. The loss function was the symmetric cross-entropy objective, as introduced in CLIP.[22] Performance during training was tracked by the median percentile rank of true report retrieval via cosine similarity for ECG-TTE pairs in the validation set.

Zero-Shot Classification

For each SHD phenotype, embeddings from positive (case) and negative (control) training cases were averaged to create centroids, defined as the arithmetic mean vector across positive (or negative) observations in the training dataset (**Fig. 2b**). For unseen data in the temporally distinct test set (2022–2023), we computed embeddings by passing the index EHR records and ECG images through the pretrained CLMBR-T and ViT models, respectively. For each test example, we then calculated cosine similarities to the two reference centroids derived from the training set: one representing known positive cases and the other known controls. The final classification score was defined as the arithmetic difference in similarity to the positive versus the negative centroid. We used this score to estimate the area under the receiver operating characteristic curve (AUROC) separately for each label. For CLMBR-T embeddings, we further performed an ablation experiment masking key demographics (sex, race, ethnicity) and extracting reference embeddings without including this key information.

Targeted vs Untargeted Screening Strategies

We operationalized two deployment paradigms for AI-ECG evaluation: an *untargeted* strategy that broadly applies AI-ECG to all individuals with available data, and a targeted strategy that prioritizes individuals based on their EHR-derived disease likelihood and ECG characteristics. In the *untargeted* approach, a logistic regression classifier was trained solely on ECG ViT embeddings and applied to every patient. In the *targeted* (sequential) strategy, individuals were first scored by a high-recall EHR model (CLMBR-T). Only those above a ≥90% sensitivity threshold proceeded to the ECG ViT classifier, while everyone else was automatically defined as negative screen. As part of the untargeted scenario, we also explored an integrated model that concatenated ECG ViT and CLMBR-T embeddings for direct classification. Thresholds for the ECG-based models were tuned on the training set via F1 optimization, whereas the EHR gating threshold was selected to achieve ≥90% recall. During testing, we applied each strategy and models directly to temporally distinct YNHHS (2022-2023) and geographically distinct UK Biobank cohorts. In the EHR, we assumed that any patient who did not pass the initial EHR CLMBR-T threshold was a predicted "negative" case, and that AI-ECG was never deployed. During external evaluation in the UK Biobank, we report metrics for the population that was selected by CLMBR-T vs the total unselected population. This is in line with the proposed, prospective implementation of the tool across community-based screening strategies.

Statistical analysis

We summarize continuous variables using medians [25th-75th percentiles], and categorical variables as counts (percentages). The discrimination performance of the different strategies for each one of the 27 SHD labels was evaluated using the area under the AUROC and derived standard errors,[23] as well as threshold-dependent metrics, such as F1-score, sensitivity, specificity, positive (PPV) and negative predictive value (NPV). All binary thresholds for classification were derived solely in the training set: F1 optimization was used to maximize precision-recall balance for the ECG ViT classifiers, while the gating threshold for CLMBR-T-based classifiers was chosen to meet a recall of ≥90%, as described above. All statistical analyses were performed in Python 3.9.

RESULTS

Data Source and Study Population

The *development cohort* consisted of 159,322 adult individuals across YNHHS (2013-2021). In total, we extracted 754,533 unique ECG-TTE pairs, in addition to a list of 118,734,437 eligible EHR-coded events from the same individuals. The training set included a balanced representation of men and women (80,275 [50.4%] female), with nearly 1 in 4 (38,062 [23.9%]) self-reporting non-White race. A separate dataset of 8,979 individuals with linked ECG-TTE studies from 2022-2023, who were *not* included in the development set, was used as a *temporally distinct test cohort*. Among these, 5,198 individuals had their first known TTE within 90 days *following* an ECG. The *geographically distinct test cohort* from the UK Biobank included 33,518 individuals who underwent protocolized ECG and CMR imaging (see **Table 1**).

Learning EHR and ECG Image Representations of SHD

Our ECG image (vision)-TTE (language) contrastive model was trained for a total of 100 epochs, with validation loss reaching a plateau (**Fig. S2**). In the validation set, the trained encoder ranked the correct TTE report at the 85.6th percentile [IQR: 64.6–96.1] for its corresponding ECG image. During temporally distinct testing (n=8,979), zero-shot classification of new cases based on the reference centroids of SHD cases and controls from the training set, successfully discriminated 27 SHD labels with a median AUROC of 0.78 [25th-75th percentile: 0.75-0.80] in the temporally distinct dataset (**Fig. 3a**), including left ventricular systolic dysfunction with an AUROC of 0.90, severe aortic stenosis with an AUROC of 0.85, elevated right ventricular systolic pressure (RVSP ≥50 mm Hg) with an AUROC of 0.82, and mitral stenosis with AUROC of 0.75. For reference, similar comparison of the associated EHR CLMBR-T embeddings at the time the ECG was recorded yielded a median AUROC of 0.62 [25th-75th perc.: 0.60-0.66]. Importantly, these metrics were comparable in an ablation study that removed sex, race and ethnicity from the EHR embeddings (median AUROC of 0.61 [25th-75th perc.: 0.69-0.65]), suggesting that the ability of EHR embeddings to discriminate SHD did not depend on these protected features (**Fig. 3b**).

Joint Modeling of EHR and ECG Image Embeddings for Targeted AI-ECG Deployment

When combined with the EHR-based CLMBR-T-extracted EHR phenotypes, this two-fold approach allowed us to define two distinct embeddings whenever an ECG is documented in the EHR; i) a 768-dimensional representation derived from the last hidden CLMBR-T layer reflecting an individual's

clinical (EHR) trajectory up until that point; and ii) a 512-dimensional representation from the projection embedding of the ECG ViT which reflects learned associations across the spectrum of echocardiographic phenotypes (**Fig. S3**). To explore the relative performance of targeted versus untargeted screening, we evaluated a pipeline that sequentially optimized each embedding (EHR CLMBR-T, and ECG ViT) for each one of the 27 reference SHD labels. We found that, compared with the untargeted approach, a "targeted" sequential strategy yielded significantly higher F1 scores (values range from 0 to 1) with a median absolute increase of 0.25 [25th-75th perc.: 0.19-0.41] across 27 SHD labels (**Fig. 4a**), largely driven by a 0.48 [25th-75th perc.: 0.40-0.65] median absolute increase in precision (PPV) (**Fig. 4b**), without notable change in recall (sensitivity) (median absolute change, 0.08 [25th-75th perc.: -0.01 to 0.22]) (**Fig. 4c**). These findings translated into an estimated 52.6% [25th-75th perc.: 46.1%-62.5%] relative reduction in the number of total positive screens (**Fig. 4d**; reflecting the number of times an echocardiogram would have been recommended) and 87.8% [25th-75th perc.: 82.4%-98.2%] relative reduction in false positive findings.

The findings were robust to a sensitivity analysis where the untargeted approach used concatenated embeddings of both the ECG ViT and EHR CLMBR-T encoders, with a median absolute increase in F1 of 0.23 [25th-75th perc.: 0.19-0.39], PPV of 0.39 [25th-75th perc.: 0.34-0.56], and sensitivity of 0.16 [25th-75th perc.: 0.09-0.27], median relative decrease of 27.4% [25th-75th perc.: 19.7%-36.6%] in AI-ECG testing, and 81.4% [25th-75th perc.: 69.7%-96.1%] relative decrease in false positive screens across the 27 SHD labels (**Fig. S4**).

External Validation

Upon external deployment in the UK Biobank (**Fig. 5a**), we first confirmed the zero-shot ability of the ECG ViT to classify structural abnormalities by CMR, including LVSD (AUROC of 0.85), increased left ventricular thickness (AUROC of 0.85), and left atrial enlargement (AUROC 0.88) (**Fig. 5b**). Across five key representative labels, we confirmed a 36.8% relative increase in F1-scores [25th-75th perc: 16.3%-54.4%] (median absolute increase of 0.02 [25th-75th perc.: 0.02-0.03]). When considering the population selected for AI-ECG screening in the targeted approach (n=9,075 individuals) versus the broader untargeted population (n=33,518 individuals), this translated into a 61.7% [25th-75th perc.: 50.4-89.1%] relative decrease in false positives (median absolute decrease of 572 [25th-75th perc.: 301-3,636]; **Fig. 5c**).

DISCUSSION

This work introduces two key innovations for the scalable and clinically meaningful deployment of AI-ECG. First, we develop and open-source an ECG image-based foundation model, by leveraging contrastive pretraining at scale across linked echocardiographic reports to learn generalizable representations spanning the full spectrum of SHD phenotypes. Second, we integrate this vision model with a longitudinal EHR foundation model, jointly modeling spot ECG-derived embeddings and dynamic patient trajectories to inform a targeted AI-ECG screening strategy. This multimodal approach enables greater precision in the identification of individuals most likely to benefit from AI-ECG evaluation, potentially reducing unnecessary deployment and improving diagnostic yield. Across temporally and geographically distinct cohorts, this framework improves the balance of precision and recall across diverse

SHD phenotypes, while substantially reducing false positives and testing burden. Together, these findings define a foundational, interoperable strategy for deploying AI-ECG in real-world care pathways, maximizing clinical utility with potential unintended downstream costs.

As the number of AI-ECG models continues to rise exponentially,[1] little guidance exists on how to implement these tools effectively within real-world clinical care pathways.[24] The controlled environment of retrospective case-control validation studies where most models are developed and tested does not replicate the multimodal reasoning that underpins clinical decision-making, which draws on both recent diagnostics and longitudinal patient history.[2] In practice, test interpretation is inherently contextual, influenced by an individual's trajectory, comorbidities, and prior risk assessments, factors that are often missing from AI deployment pipelines. To address this gap, we introduce an integrated strategy that combines two complementary foundation models: an ECG image-based vision transformer, designed for direct and flexible use with raw ECG images (e.g., screenshots or smartphone captures),[16–19] and a pre-trained CLMBR-T transformer that encodes longitudinal EHR trajectories using standardized, interoperable data schemas.[13,14] While prior models have largely relied on static risk factors or subjective subgroup definitions to define at-risk populations, our approach enables a dynamic, data-driven gating system that selects candidates for AI-ECG screening based on their evolving clinical context in the EHR. Importantly, we show that neither modality alone is sufficient: EHR embeddings lack the phenotypic specificity captured by ECG signals, while ECG-only models may lack awareness on key features that best define an individual's pre-test probability and thus the optimal interpretation of a positive versus negative screen. Instead, we find that their joint integration allows for high-fidelity, high-precision screening. Given an estimated cost of ~$125 per AI-ECG-enabled interpretation per the 2025 Centers for Medicare & Medicaid Services, this potential reduction in low-yield model deployment has immediate implications for scalability and cost-effectiveness across health systems.

Our approach has key strengths with direct clinical implications. First, it introduces the first standalone AI-ECG foundation model for SHD screening using ECG images, the most accessible, interoperable, and clinically ubiquitous format, enabling scalable diagnostic deployment across care settings, including remote and resource-limited environments. Second, it establishes and validates the novel concept that longitudinal EHR embeddings can function as interoperable gatekeepers for AI deployment, identifying individuals most likely to benefit from AI diagnostics. This enables precise targeting of AI-ECG, reducing unnecessary model usage and false positives, two critical components for real-world adoption at scale. Given its foundational nature, the system also supports dynamic retuning to accommodate new disease labels and evolving practice patterns.[25] Importantly, the EHR-derived risk enrichment operates largely independently of protected demographic variables such as sex, race, or ethnicity, promoting equity in screening.[26,27] Third, the framework is built with a high degree of interoperability across both ECG and EHR domains, covering a broad spectrum of SHD phenotypes and supporting integration into heterogeneous health systems. Finally, by open-sourcing the model weights and codebase, we empower both individual investigators and institutions to fine-tune and adapt the system to their unique populations, accelerating innovation and deployment. The underlying training cohort is large and demographically diverse, providing a robust foundation for domain-specific extensions and future discovery.

LIMITATIONS

Certain limitations merit consideration. First, the EHR-derived embeddings reflect longitudinal trajectories of structured phenotypes, and do not account for unstructured information in the form of encounter notes. Second, although the UK Biobank provides insights into the external reproducibility of our approach, systematic differences in the design of this cohort may limit the interpretation of our findings. Unlike the EHR where clinical tests are part of clinical care, the UK Biobank involved protocolized measurement of blood/serum biomarkers. Similarly, the decision to pursue ECG and CMR imaging was based on a research protocol without confounding by indication. Finally, prospective studies should explore the passive and active deployment of this targeted screening strategy across the EHR of large health systems to examine its validity and robustness to potential data drifts or shifts.

CONCLUSION

We introduce *TARGET-AI*, a foundational strategy for the targeted deployment of AI-enabled screening technologies in the EHR. By selectively enriching the target populations through multi-modal integration, our platform paves the path towards an interoperable strategy for the optimal deployment of AI-ECG or any AI-assisted technology, which can be flexibly adapted across health systems and for any computable label.

DATA AVAILABILITY:

The sample code for training and inference, including a pipeline for embedding extraction from new ECG images, as well as reference embeddings for case and control centroids across 27 representative echocardiographic labels from the YNHHS training set, will be released through our Lab's GitHub repository upon formal acceptance of the peer-reviewed manuscript. Model weights for the ECG image vision transformer will also be made available through our Lab's HuggingFace repository.

DISCLOSURES:

E.K.O. is an Associate Editor for European Heart Journal, a co-founder of Evidence2Health LLC, a co-inventor in patent applications 18/813,882, 17/720,068, 63/508,315, 63/580,137, 63/619,241, 63/562,335, and granted patents US12067714B2, US11948230B2, has been an ad hoc consultant for Caristo Diagnostics Ltd and Ensight-AI Inc, and has received royalty fees from technology licensed through the University of Oxford. R.K. is an Associate Editor of JAMA and receives research support, through Yale, from the Blavatnik Foundation, Bristol-Myers Squibb, Novo Nordisk, and BridgeBio. He is a coinventor of Pending Patent Applications WO2023230345A1, US20220336048A1, 63/346,610, 63/484,426, 63/508,315, 63/580,137, 63/606,203, 63/619,241, and 63/562,335, and a co-founder of Ensight-AI, Inc and Evidence2Health, LLC. The remaining authors have nothing to disclose.

REFERENCES:

1. Siontis, K. C., Noseworthy, P. A., Attia, Z. I. & Friedman, P. A. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat. Rev. Cardiol.* **18**, 465–478 (2021).
2. Khera, R. *et al.* Transforming cardiovascular care with artificial intelligence: From discovery to practice: JACC state-of-the-art review. *J. Am. Coll. Cardiol.* **84**, 97–114 (2024).
3. Rubin, R. Obstacles to implementing AI tools in health care. *JAMA* **325**, 333 (2021).
4. Han, R. *et al.* Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit. Health* **6**, e367–e373 (2024).
5. Johnson, L. S. *et al.* Artificial intelligence for direct-to-physician reporting of ambulatory electrocardiography. *Nat. Med.* 1–7 (2025).
6. Jain, S. S., Mello, M. M. & Shah, N. H. Avoiding financial toxicity for patients from clinicians' use of AI. *N. Engl. J. Med.* (2024) doi:10.1056/nejmp2406135.
7. Elvidge, J. & Dawoud, D. Reporting standards to support cost-effectiveness evaluations of AI-driven health care. *Lancet Digit. Health* **6**, e602–e603 (2024).
8. Badal, K., Lee, C. M. & Esserman, L. J. Guiding principles for the responsible development of artificial intelligence tools for healthcare. *Commun. Med. (Lond.)* **3**, 47 (2023).
9. Lekadir, K. *et al.* FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **388**, e081554 (2025).
10. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
11. Vaid, A. *et al.* A foundational vision transformer improves diagnostic performance for electrocardiograms. *NPJ Digit Med* **6**, 108 (2023).
12. Christensen, M., Vukadinovic, M., Yuan, N. & Ouyang, D. Vision-language foundation model for echocardiogram interpretation. *Nat. Med.* **30**, 1481–1488 (2024).
13. Wornow, M., Thapa, R., Steinberg, E., Fries, J. A. & Shah, N. H. EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. *arXiv [cs.LG]* (2023).
14. Steinberg, E. *et al.* Language models are an effective representation learning technique for electronic health record data. *J. Biomed. Inform.* **113**, 103637 (2021).
15. Bai, W. *et al.* A population-based phenome-wide association study of cardiac and aortic structure and function. *Nat. Med.* **26**, 1654–1662 (2020).
16. Sangha, V. *et al.* Automated multilabel diagnosis on electrocardiographic images and signals. *Nat. Commun.* **13**, 1583 (2022).
17. Sangha, V. *et al.* Detection of Left Ventricular Systolic Dysfunction From Electrocardiographic Images. *Circulation* (2023) doi:10.1161/CIRCULATIONAHA.122.062646.
18. Oikonomou, E. K. *et al.* Artificial Intelligence-Enhanced Risk Stratification of Cancer Therapeutics-Related Cardiac Dysfunction Using Electrocardiographic Images. *Circ: Cardiovascular Quality and Outcomes* (2024) doi:10.1161/CIRCOUTCOMES.124.011504.
19. Dhingra, L. S. *et al.* Heart failure risk stratification using artificial intelligence applied to electrocardiogram images: a multinational study. *Eur. Heart J.* (2025) doi:10.1093/eurheartj/ehae914.

20. Holste, G., Oikonomou, E. K., Wang, Z. & Khera, R. PanEcho: Complete AI-enabled echocardiography interpretation with multi-task deep learning. *medRxiv* 2024.11.16.24317431 (2024) doi:10.1101/2024.11.16.24317431.
21. Bao, H., Dong, L., Piao, S. & Wei, F. BEiT: BERT Pre-Training of Image Transformers. *arXiv [cs.CV]* (2021).
22. Radford, A. *et al.* Learning transferable visual models from natural language supervision. *arXiv [cs.CV]* (2021).
23. Hanley, J. A. & McNeil, B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**, 839–843 (1983).
24. Longhurst, C. A., Singh, K., Chopra, A., Atreja, A. & Brownstein, J. S. A call for artificial intelligence implementation science centers to evaluate clinical effectiveness. *NEJM AI* **1**, (2024).
25. Youssef, A. *et al.* External validation of AI models in health should be replaced with recurring local validation. *Nat. Med.* **29**, 2686–2687 (2023).
26. Ratwani, R. M., Sutton, K. & Galarraga, J. E. Addressing AI algorithmic bias in health care. *JAMA* **332**, 1051–1052 (2024).
27. Hswen, Y. & Voelker, R. New AI Tools Must Have Health Equity in Their DNA. *JAMA* **330**, 1604–1607 (2023).

Figures

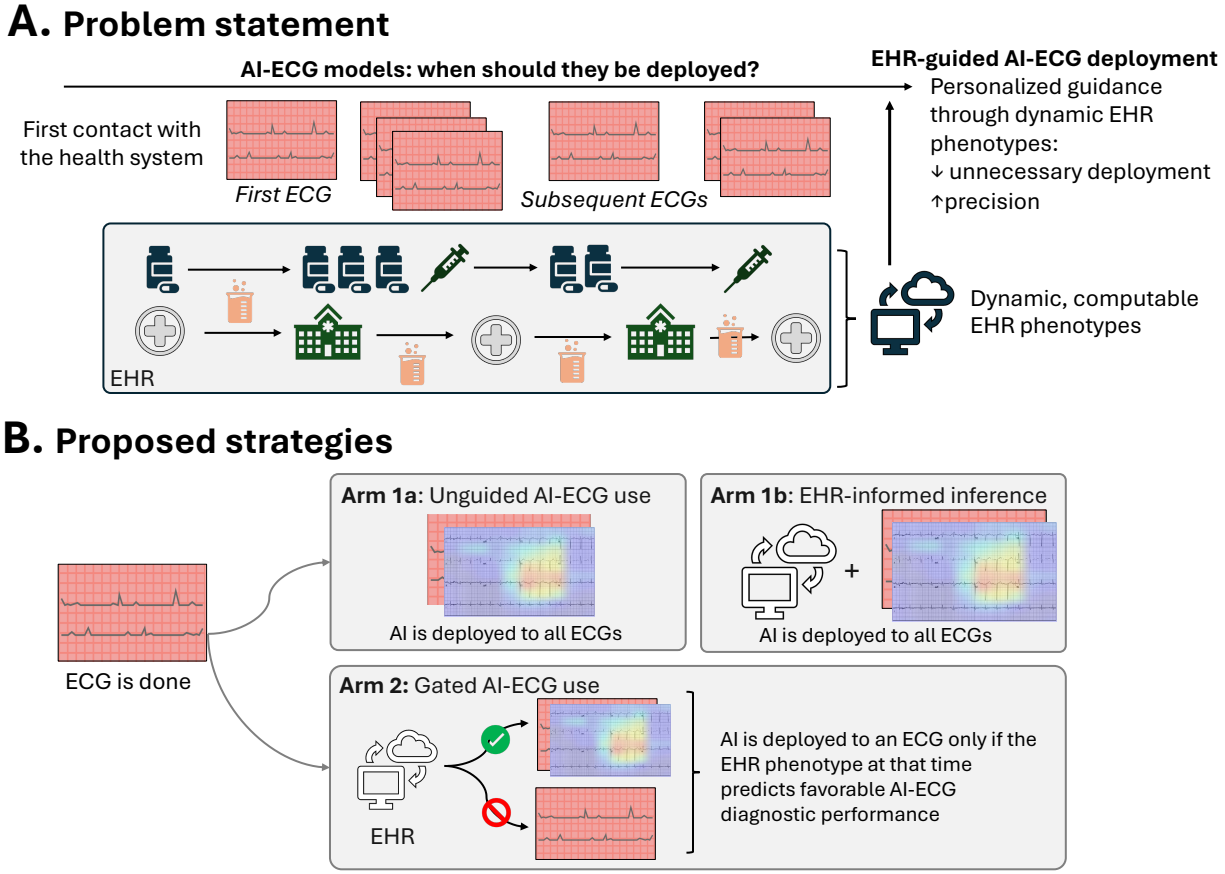

Fig. 1 | Targeted deployment of AI screening technologies in the EHR. (A) Problem statement: Across their longitudinal care in an integrated health system, individuals may receive multiple diagnostic tests, such as electrocardiograms (ECG). ECG phenotypes evolve in tandem with the broader clinical profile of an individual, as recorded and computed in the electronic health record (EHR). Artificial intelligence (AI)-guided technologies may screen for structural heart disease (SHD) directly from ECGs, but there is currently no data-driven strategy to enable their targeted deployment, to minimize false discovery and unnecessary downstream testing. **(B) Potential strategies:** The integration of AI-ECG technologies may follow different strategies. For example, AI-ECG can be deployed opportunistically across all eligible ECGs with or without incorporating EHR-derived clinical characteristics (*"Untargeted Approach" - Arm 1*). Alternatively, the computable phenotypes of individuals in the EHR may be leveraged as gatekeepers to prevent unnecessary deployment of AI-ECG in cases, optimizing the precision of the technology ("*Targeted Approach*" - *Arm 2*). *AI: artificial intelligence; ECG: electrocardiography; EHR: electronic health record.*

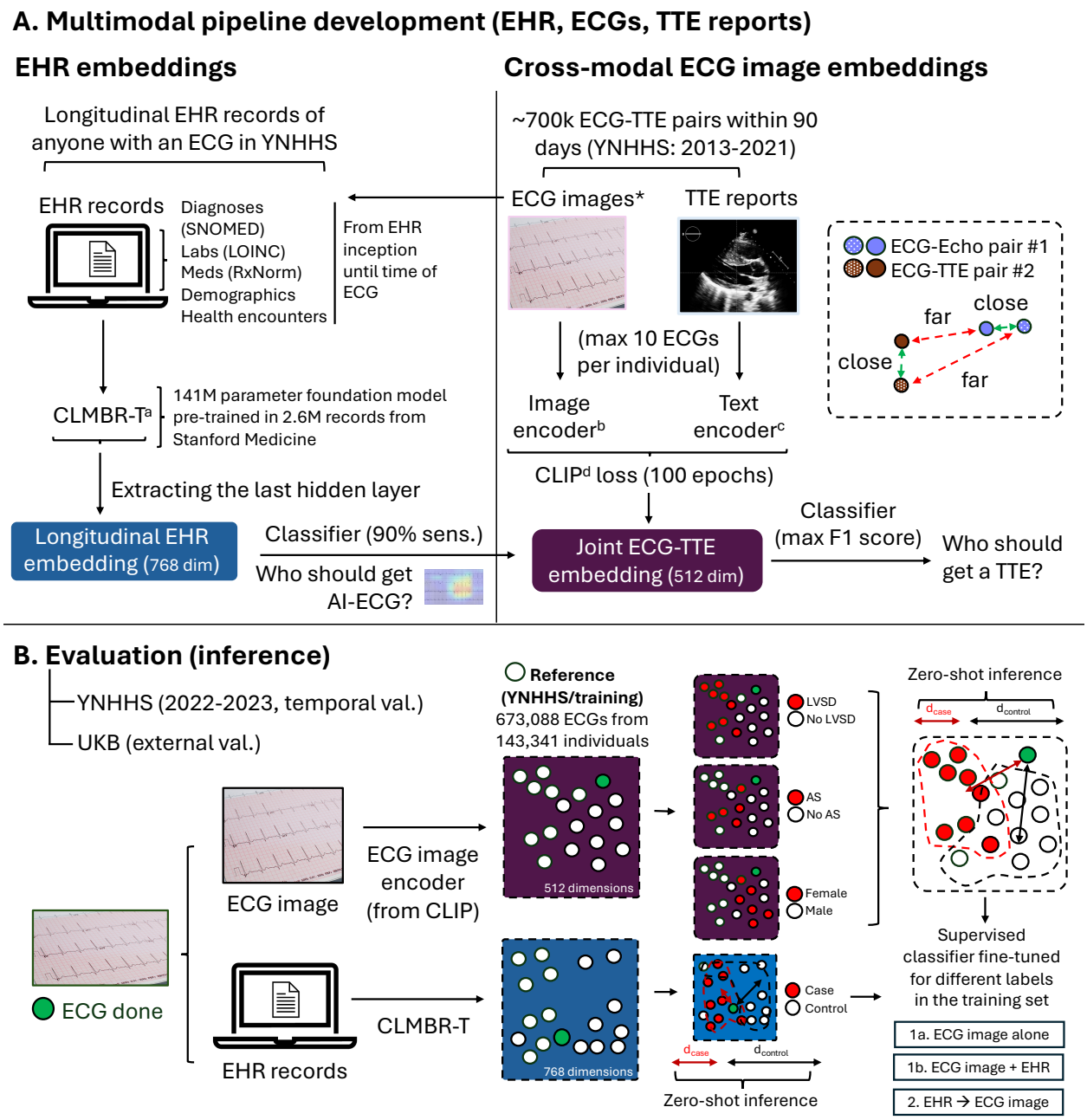

Fig. 2 | Study design. (A) Training: To evaluate the feasibility and performance of leveraging the EHR for the targeted screening of structural heart disease (SHD) on electrocardiography (ECG), we designed a strategy that jointly leverages two foundation models. First, a pre-trained foundation model built for longitudinal EHR phenotypes (CLMBR-T) is used to define an embedding of an individual's longitudinal EHR trajectory up until the time an ECG is performed (spanning diagnosis codes, concepts, medications, laboratory measurements, visits). Second, we apply contrastive pre-training to learn a joint embedding between ECG images (using a vision transformer) and text embeddings of corresponding transthoracic echocardiography reports performed within 90 days of each other in the YNHHS cohorts (2016-2021). This allows us to extract a 768-dimensional embedding reflecting an individual's EHR history up until that point, as well as a 512-dimensional embedding reflecting structural correlates from the ECG (image).

For each one of 27representative echocardiographic SHD labels, covering ventricular, atrial, and valvular function and anatomy, we first trained label-specific classifiers on EHR embeddings to achieve 90% sensitivity. We then trained separate classifiers on ECG embeddings, restricted to the subset of patients who met the EHR-based threshold. **(B) Evaluation:** The models were evaluated in a temporally distinct dataset of individuals undergoing ECG at YNHHS between 2022 and 2023 who were not seen during training and who had not undergone an echocardiogram in our system prior to the index ECG, as well as in individuals in the UK Biobank who enrolled in the ECG and concurrent CMR assessment. We evaluate both the zero-shot performance of the embeddings in identifying discrete anomalies by leveraging the relative cosine similarity of a new case versus the reference embeddings of individuals with and without the condition of interest in the training set. Finally, we evaluate the performance of a targeted vs untargeted AI-ECG screening strategy using the fine-tuned sequential "*targeted*" (EHR pre-screening, followed by AI-ECG) or "*untargeted*" models (AI-ECG for all). *AI: artificial intelligence; CLIP: Contrastive Language-Image Pre-training; CLMBR: clinical language model-based representation - transformer; CMR: cardiac magnetic resonance imaging; ECG: electrocardiography; EHR: electronic health record; TTE: transthoracic echocardiography; YNHHS: Yale-New Haven Health System.*

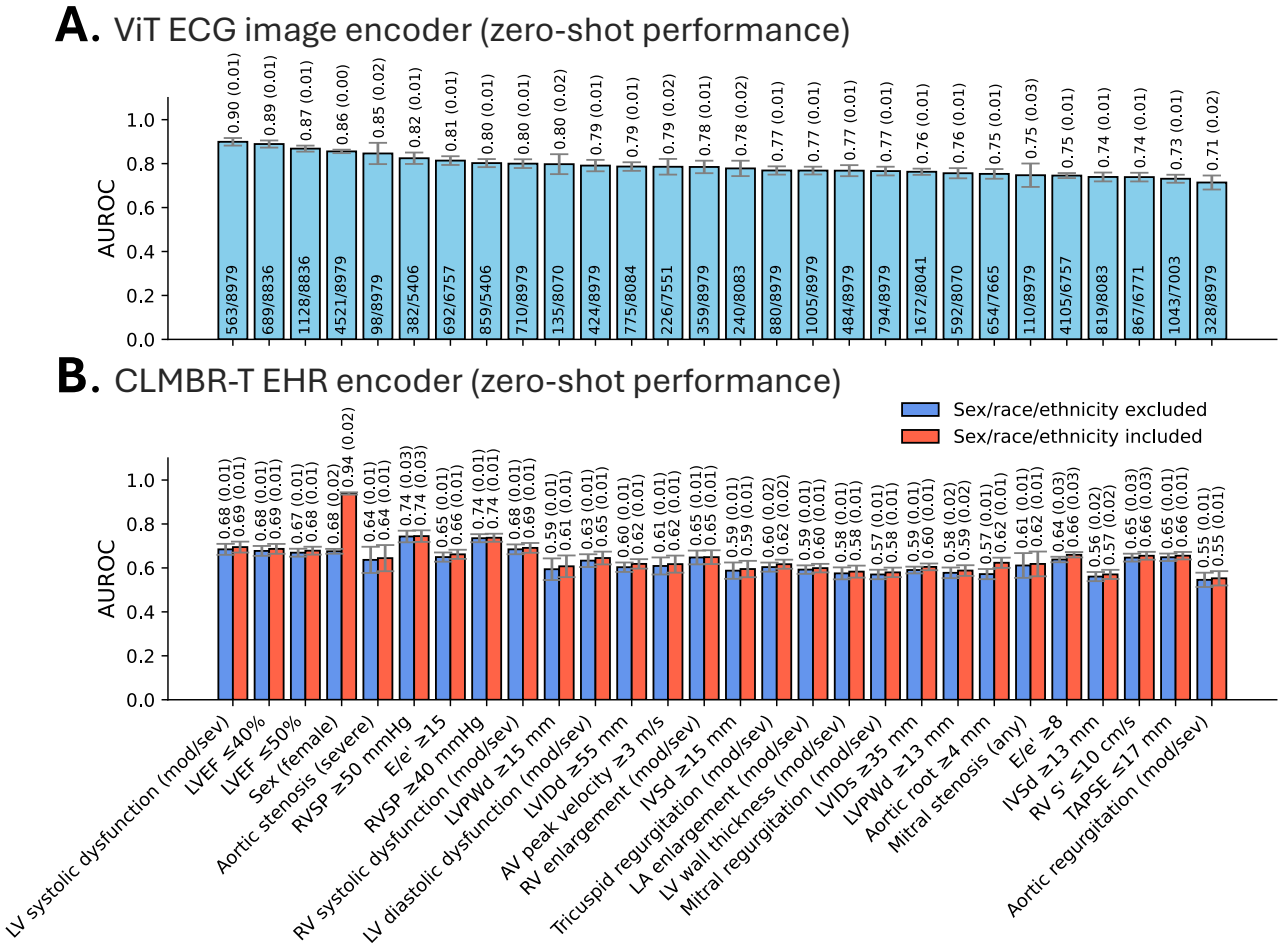

Fig. 3 | Zero-shot evaluation of AI-ECG ViT and CLMBR-T across the spectrum of echocardiographic phenotypes. During zero shot evaluation, we use the ECG Image ViT and EHR CLMBR-T to project each new case to the respective embeddings. For each structural heart disease (SHD) label (ascertained based on an echocardiogram performed within 90 days), we calculate the absolute difference of the cosine similarity to the centroid of all observations in the training set that were cases versus the centroid of all controls. This relative similarity is then used to rank the test set observations for a given label and compute discrimination metrics (AUROC), which are then visualized using bar plots for the ViT ECG Image **(A)** and CLMBR-T encoders **(B)**, respectively. We present two versions of CLMBR-T metrics, both without (red) and with (blue) protected demographics (sex, race, ethnicity). Error bars denote 95% confidence intervals. *AI: artificial intelligence; AUROC: area under the receiver operating characteristic curve; AV: aortic valve; E/e': ratio of early mitral inflow velocity to early diastolic mitral annular velocity; ECG: electrocardiography; IVSd: interventricular septum thickness at end-diastole; LA: left atrium; LV: left ventricle; LVEF: left ventricular ejection fraction; LVIDd: left ventricular internal diameter at end-diastole; LVIDs: left ventricular internal diameter at end-systole; LVPWd: left ventricular posterior wall thickness at end-diastole; RA: right atrium; RV: right ventricle; RVSP: right ventricular systolic pressure; TAPSE: tricuspid annular plane systolic excursion; ViT: vision transformer.*

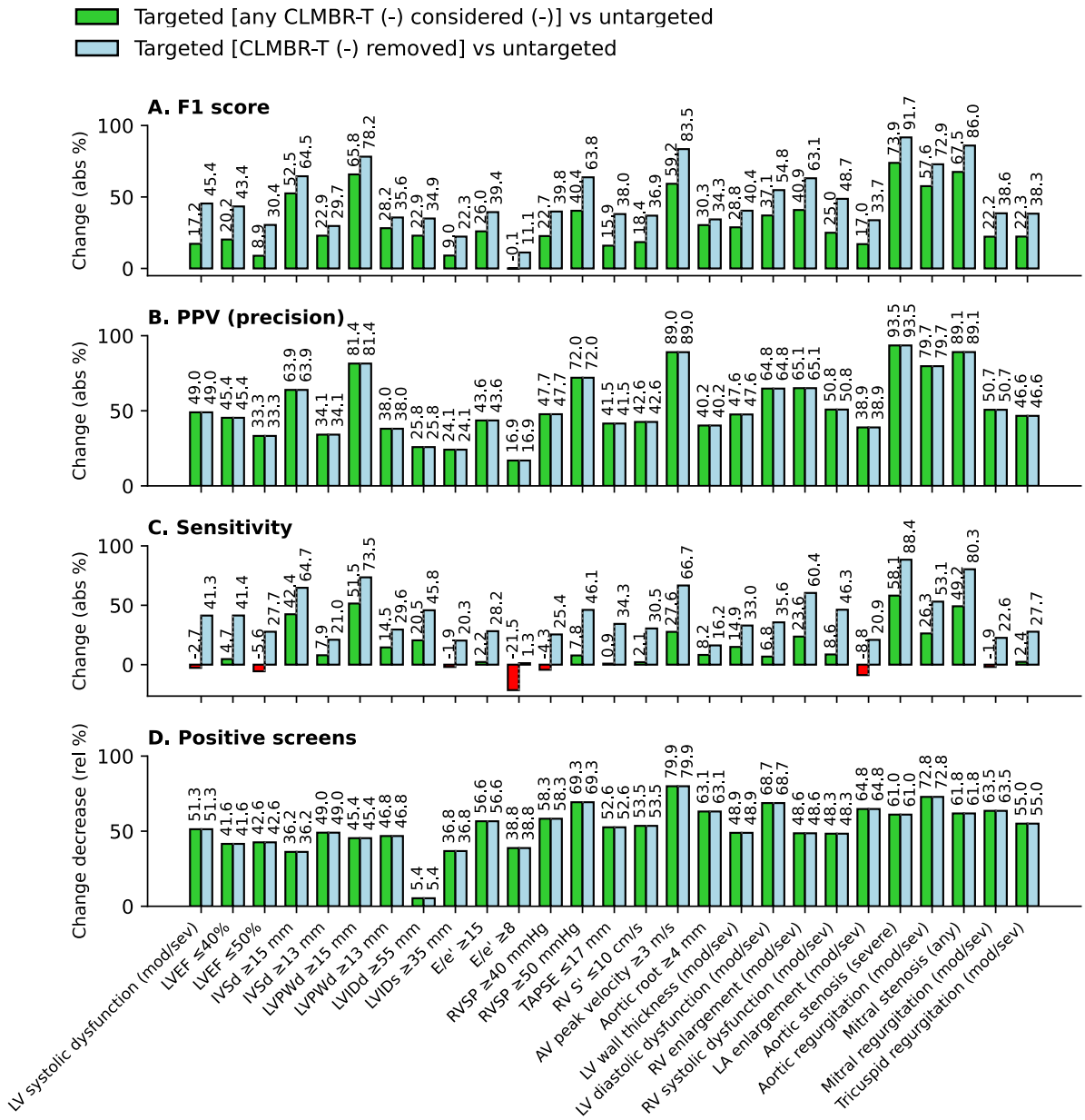

Fig. 4 | Comparison of untargeted vs targeted AI-ECG screening for structural heart disease. We summarize the absolute (or relative) difference in metrics of discrimination for various forms of structural heart disease (as defined by transthoracic echocardiography) for a targeted (CLMBR-T-guided EHR pre-screening, followed by targeted AI-ECG) vs untargeted AI-ECG approach (AI-ECG applied in all). Green color bars present data in the overall population, classifying any individuals not meeting the initial CLMBR-T-defined bar as negative for a given condition. The blue color bars present results only for the valid population that passed the initial CLMBR-T threshold. (**A**) Absolute change in F1 score; (**B**) Absolute change in positive predictive value (PPV); (**C**) Absolute change in sensitivity; (**D**) Relative decrease in total positive screens. *AI: artificial intelligence; AUROC: area under the receiver operating*

characteristic curve; AV: aortic valve; E/e': ratio of early mitral inflow velocity to early diastolic mitral annular velocity; ECG: electrocardiography; IVSd: interventricular septum thickness at end-diastole; LA: left atrium; LV: left ventricle; LVEF: left ventricular ejection fraction; LVIDd: left ventricular internal diameter at end-diastole; LVIDs: left ventricular internal diameter at end-systole; LVPWd: left ventricular posterior wall thickness at end-diastole; RA: right atrium; RV: right ventricle; RVSP: right ventricular systolic pressure; TAPSE: tricuspid annular plane systolic excursion; ViT: vision transformer.

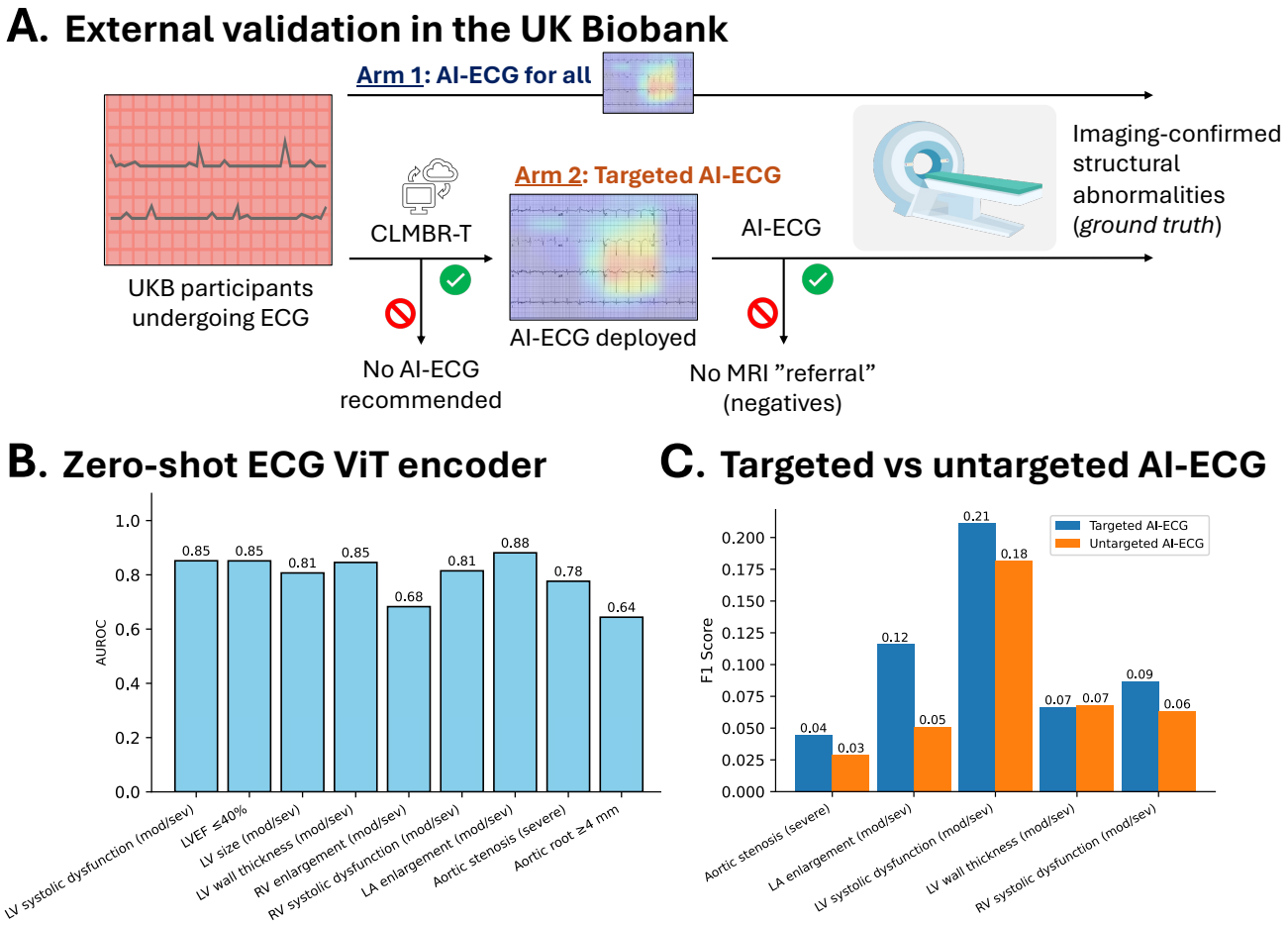

Fig. 5 | External validation in the UK Biobank. **(A)** External validation of the targeted AI-ECG approach in the UK Biobank among 33,518 UK Biobank participants undergoing concurrent research ECG and CMR imaging. We applied the pre-trained encoders and label-specific supervised headers in parallel for both the untargeted and the targeted approach, with structural heart disease defined based by CMR. **(B)** Zero-shot performance of the ECG image ViT for structural heart disease. For size and thickness measurements, we present discrimination for individuals falling in the ≥99th percentile of the respective distribution (counts reflect positive labels and total valid observations for a given label). Aortic stenosis was adjudicated based on ICD-10 diagnosis codes (e.g., I35-) present on or before the date of CMR **(C)** Summary of F1-score for the targeted vs untargeted AI-ECG screening strategy across eligible labels. AI: artificial intelligence; CMR: cardiac magnetic resonance; ECG: electrocardiography; MRI: magnetic resonance imaging.

Tables

Table 1 | Study Population Demographics & Clinical Characteristics.

	YNHHS		UK Biobank
	Development	**Testing**	
Year of Visit	2013 to 2021	2022 or 2023	2014-2020
Number of ECGs	754,533	8,979	33,518
Number of unique participants	159,322	8,979	33,518
Age at ECG (years)	68 [57-78]	68 [58-78]	65 [58-70]
Sex			
Female	80,275 (50.4)	4,521 (50.4)	17,463 (52.1)
Male	79,034 (49.6)	4,458 (49.6)	16,055 (47.9)
Not specified/Missing	13 (0.0)	-	-
Race			
American Indian or Alaska Native	479 (0.3)	28 (0.3)	0 (0)
Asian	2,764 (1.7)	183 (2.0)	445 (1.3)
Black or African American	20,531 (12.9)	1,161 (12.9)	219 (0.7)
Native Hawaiian or Other Pacific Islander	313 (0.2)	18 (0.2)	0 (0)
White	121,260 (76.1)	6,723 (74.9)	32,430 (96.8)
Other/Unknown	13,975 (8.8)	866 (9.6)	424 (1.3)
Ethnicity			
Hispanic	14,034 (8.8)	902 (10.0)	-
Non-Hispanic	140,977 (88.7)	7,801 (86.9)	-
Unknown	4,311 (2.5)	276 (3.1)	-
Extractable computable phenotypes*			
Hypertension	117,398 (73.7)	6,485 (72.2)	6,331 (14.8)
Diabetes mellitus	42,735 (26.8)	2,272 (25.3)	1,289 (3.0)
Hypercholesterolemia	34,930 (21.9)	1,917 (21.4)	2,779 (6.5)
Chronic cardiac disease	73,225 (46.0)	3,243 (36.1)	2,457 (5.8)
Heart failure	37,710 (23.7)	1,689 (18.8)	228 (0.5)
Stroke	33,275 (20.9)	1,646 (18.3)	585 (1.4)
Chronic kidney disease	15,886 (10.0)	551 (6.1)	127 (0.3)
Chronic respiratory disease	55,904 (35.1)	2,801 (31.2)	1,069 (2.5)
Cancer	20,159 (12.7)	1,126 (12.5)	2,991 (7.0)

Abbreviations: ECG: electrocardiography; EHR: electronic health record; YNHHS: Yale-New Haven Health System. Data are presented as counts (percentages) or median [25th-75th percentile], as appropriate. *inpatient-linked encounters in the UK Biobank.