Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Green means go: an ECG foundation model over-calls normal athletes, and athlete-criteria labels correct it

Chukwudi Eke

chukwudieke61@gmail.com

Independent Researcher https://orcid.org/0009-0002-2294-673X

Research Article

Keywords: electrocardiography, athletes, screening, artificial intelligence, sports cardiology

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11776488/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Objective

To test whether a publicly released ECG foundation model respects the International Recommendations for ECG interpretation in athletes, and whether its errors can be corrected without athlete data.

Methods

We applied ECGFounder, trained on about 10 million hospital ECGs, to non-athletes with only training-related (“green”) findings (PTB-XL, Chapman/Ningbo) and to two athlete groups: 26 Norwegian endurance athletes read as normal by a cardiologist and 156 green-only recordings from 52 Spanish footballers. In a sequence of pre-registered steps we tested four fixes, anchoring sensitivity on screening-relevant abnormal ECGs, and confirmed the best fix with thresholds set on data it had never seen.

Results

Green-only ECGs were called abnormal in 33.7% of PTB-XL and 17.9% of Chapman/Ningbo cases, and the model called abnormal 34.6% of normal endurance athletes and 47.4% of footballer recordings, sometimes labelling infarction. Rules applied over the model’s own named findings, with or without measured intervals, made things worse. A small readout retrained on the model’s features with athlete-criteria labels, using non-athletes only, called 17.6% (95% CI 10.3% to 25.4%) of normal athletes abnormal against 52.8% for the model’s own output, at 87.5% against 89.4% sensitivity for screening-relevant abnormal ECGs, and held across three training seeds.

Conclusion

The model’s representation already separates athletic adaptation from disease; its default output does not. Relabelling the readout against the International Criteria removes about two-thirds of the false alarms.

Summary box

WHAT IS ALREADY KNOWN ON THIS TOPIC

- Training changes athletes’ ECGs in ways that are normal for them, and the International Recommendations list these “green” findings, which need no further testing.
- Before athlete-specific criteria, general-population rules sent many healthy athletes for unnecessary tests.

- ECG foundation models are trained on hospital ECGs described with general-population conventions.

WHAT THIS STUDY ADDS

- A publicly released ECG foundation model called a third to a half of normal athletes abnormal, at times labelling infarction, and a slow heart rate below 50 beats per minute drove most of it.
- A readout retrained with athlete-criteria labels on non-athletes cut false alarms in normal athletes from 53% to 18% at almost unchanged sensitivity, whereas rules over the model's own findings made things worse.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- ECG foundation models should not be used in athlete screening with their default outputs.
- Readouts labelled against the International Criteria are a cheap correction, and they should be validated prospectively in larger and more diverse athlete groups, including Black athletes and women.

Introduction

Many sports organisations screen athletes with an ECG before they compete, in a bid to find the rare heart conditions that can cause sudden death during exercise. Regular intense training, however, changes the heart's structure and electrical activity, and while these changes are healthy, several of them look like disease when judged by general-population rules. Examples include a slow resting heart rate (sinus bradycardia), a slightly delayed conduction between the upper and lower chambers (first-degree atrioventricular block), an incomplete right bundle branch block, tall QRS voltages that meet voltage criteria for left ventricular hypertrophy, and early repolarisation.

Early screening programmes applied general rules to athletes, and many healthy athletes were sent for unnecessary tests or withheld from sport[1–3]. The International Recommendations for ECG interpretation in athletes were written to stop this[4], and they divide findings into **green** training-related findings that need no further evaluation, **borderline** findings that matter only in combination, and **abnormal** findings that need further testing. The recommendations are being revised again, as the 2025 International Summit has proposed changes to the rules for T-wave inversion and premature ventricular contractions, while leaving the green findings studied here unchanged in the published summary[5].

ECG foundation models are trained on millions of hospital ECGs and the diagnostic statements written about them[6], and those statements follow general-population conventions, so that a heart rate of 42 beats per minute is written as "marked sinus bradycardia", an incomplete right bundle branch block is filed under conduction disturbances, and high voltage is filed under hypertrophy. A model that learns these conventions faithfully will reproduce exactly the false alarms that the athlete criteria were written to remove, at scale and with an appearance of objectivity. Open ECG databases from athletes were

released specifically to support artificial intelligence research[7,8], and a recent multicentre study found that a general-purpose large language model misread athletes' ECGs, mostly by missing abnormalities[9]; to our knowledge, however, no study has tested whether an ECG foundation model respects the athlete criteria, or whether it can be made to.

We asked two questions:

1. Does a publicly released ECG foundation model call green findings abnormal, first in non-athletes who carry them and then in real athletes?
2. Can the error be corrected without athlete data, and at what cost in sensitivity for the abnormal findings that screening exists to find?

Methods

Model and data

ECGFounder is a 12-lead ECG model trained on 10.8 million ECGs from the Harvard–Emory ECG Database[6]. It outputs probabilities for 150 named diagnostic statements and an overall "abnormal ECG" probability, and none of the four datasets we used is part of its training data. We used the released weights and the published preprocessing unchanged (band-pass filter, scaling to zero mean and unit variance, 10 seconds at 500 Hz).

The non-athlete data were PTB-XL[10] and Chapman/Ningbo[11,12]. An ECG counted as *green-only* if every label given by the dataset was a green finding, sinus rhythm or "normal", with at least one green finding present (PTB-XL codes SR, NORM, SBRAD, SARRH, 1AVB, IRBBB, VCLVH and HVOLT; Chapman/Ningbo codes SR, SB, SA, 1AVB, IRBBB, ERV, JEB and WAVN). Labels came from the datasets, never from the model.

The Norwegian Endurance Athlete ECG Database[7] holds 28 elite endurance athletes recorded on a GE MAC VUE 360, each read by a cardiologist trained in athlete ECG interpretation and by the device's own algorithm (GE 12SL); our group is the 26 athletes the cardiologist read as normal. PF12RED[8] holds 163 recordings from 54 professional footballers of a Spanish first-division team. Before reading the recordings we fixed "green-only" as no inverted T wave in leads I, II, aVL, aVF or V3–V6, and we excluded the 2 players with anterior T-wave inversion, because the repolarisation variant seen in Black athletes cannot be checked with these labels, which left 52 players and 156 recordings. Rates for footballers are per recording, with confidence intervals from a bootstrap that resamples players.

Pre-registered steps

The work ran in four pre-registered steps, each written before its data were scored, and we report every hypothesis, including those that failed. Step 1 measured over-call at a threshold that flagged 10% of normal ECGs in each dataset, with the athletes' threshold (0.9845) set on 2,000 PTB-XL normal ECGs

and frozen; the hypothesis was that the model flags at least 20% of normal athletes. Step 2 tested two cheap fixes, a retrained linear readout and the athlete criteria applied over the model's 150 named findings, with sensitivity anchored on any non-green label. Step 3 re-anchored sensitivity on screening-relevant abnormal ECGs and compared four fixes along their full operating curves. Step 4 ("3b") was the confirmatory test of the best fix from step 3. Thresholds for every step are listed in online supplemental table S1.

Screening-relevant abnormal ECGs

Pre-participation screening exists to find a short list of conditions, so in step 3 we defined a screening-relevant abnormal ("screening-red") set from dataset labels: pre-excitation, long QT, complete left bundle branch block or intraventricular conduction delay, T-wave inversion and ischaemic-type repolarisation, ST depression, left ventricular hypertrophy other than voltage alone, and Q-wave infarction as a proxy for pathological Q waves (PTB-XL codes WPW, LNGQT, CLBBB, IVCD, INVT, ISC_ family, STD_, LVH and the Q-wave infarction codes; Chapman/Ningbo by the equivalent SNOMED codes). Sensitivity was measured on screening-red ECGs, false alarms on normal and green-only ECGs, and other abnormal ECGs were left out of both.

Fixes compared in step 3

All fixes used the same frozen forward pass. The *criteria rule over all findings* (step 2's rule) calls an ECG abnormal if any of about 100 red findings, or two borderline findings, exceed one shared threshold. The *screening-restricted rule* uses only the model's findings that correspond to screening-red, each with its own threshold set on normal and green-only training ECGs. The *measured-intervals hybrid* applies the International Criteria to heart rate, PR, QRS and corrected QT measured with an open delineator (NeuroKit2), and asks the model only about morphology. The two *trained readouts* are a logistic layer and a small neural network (one hidden layer of 256 units) trained on the model's internal features to separate screening-red from normal and green-only ECGs, using non-athletes only.

Confirmatory test (step 3b)

The readout was retrained on PTB-XL folds 1–6, stopped early on fold 7, and its threshold for 90% screening-red sensitivity was set on fold 8, which it had never seen; the model's own output was thresholded on fold 8 in the same way. The pre-registered hypothesis was that, against the model's own output, the readout at least halves the false-alarm rate in normal athletes (pooled, with the 95% bootstrap interval of the difference below zero) while losing no more than 2 percentage points of screening-red sensitivity on PTB-XL folds 9–10, with at least 80% sensitivity on Chapman/Ningbo. The readout was trained with three seeds, and the first was the primary test.

Statistics and reporting

Proportions carry Wilson 95% intervals, and athlete rates carry athlete-level bootstrap intervals (4,000 resamples). Rates for findings present in fewer than 20 ECGs are shown but not interpreted, and

reporting follows the applicable items of TRIPOD + AI for the evaluation of an existing model[13].

Results

Non-athletes and athletes: green findings trigger false alarms

At a threshold that flagged 10% of normal ECGs, ECGs with only green findings were flagged far more often (Fig. 1), at 33.7% (31.5 to 35.8) in PTB-XL, 3.4 times the normal rate, and 17.9% (16.3 to 19.7) in Chapman/Ningbo, 1.8 times the normal rate, and sinus bradycardia alone was flagged in 71.6% of PTB-XL cases (online supplemental table S2). At the frozen athlete threshold, the model called abnormal 34.6% (19.4 to 53.8) of the 26 endurance athletes a cardiologist had read as normal and 47.4% (36.4 to 58.5) of the 156 green-only footballer recordings, so the pre-registered hypothesis was met in both groups, although the lower end of the interval for the endurance athletes (19.4%) touches the 20% bar.

What the model said these normal athletes had is shown in Fig. 2. For the endurance athletes it reported sinus bradycardia (7), marked sinus bradycardia (6), right axis deviation (2), left atrial enlargement (2) and anterior infarction (1), and for the footballer recordings sinus bradycardia (61), marked sinus bradycardia (50), left ventricular hypertrophy (14), lateral infarction (11), right bundle branch block (10) and incomplete right bundle branch block (8). The device's own algorithm (GE 12SL), run on the same recordings, labelled 13 of the 26 normal endurance athletes borderline (7) or abnormal (6), so the foundation model reproduces the general-population convention that conventional machine reading already follows. Among green-only footballers, false alarms were 60% in players of African ancestry (5 players), 48.8% in Caucasian (40) and 23.5% in Latin American players (6), although these groups are too small for any conclusion.

Heart rate behaved as a cliff rather than a slope. Among normal and green-only non-athlete ECGs, at the step-3 operating point for 90% screening-red sensitivity, the model flagged 94.4% of PTB-XL ECGs with a heart rate below 50 beats per minute against 14.0% to 19.5% at 50 and above, and 92.4% against 8.4% to 11.3% in Chapman/Ningbo, while green findings still raised the odds of a flag 4.2-fold (PTB-XL) and 2.2-fold (Chapman/Ningbo) after allowing for heart rate. Elite athletes sit right at this edge, as the median measured heart rate was 55.8 beats per minute in the endurance athletes and 53.5 in the footballers.

Figure 1.

Share of ECGs called abnormal at an operating point that flags 10% of normal ECGs (dashed line). Blue: non-athletes with green findings only. Orange: normal athletes. Bars are 95% intervals (Wilson; footballers by player-level bootstrap).

Figure 2.

Abnormal findings the model assigned to normal athletes. Infarction labels are shown in red.

Rules over the model's findings do not fix it

The two step-2 fixes failed their pre-registered criteria. When sensitivity was anchored on any non-green label, the retrained linear readout lowered false alarms on green-only PTB-XL ECGs from 46.5% to 33.3% only by doubling them on normal ECGs, from 18.0% to 35.7%, and the criteria rule over all findings was worse than the model's own output on every false-alarm measure. Re-anchoring on screening-red ECGs in step 3 did not rescue the rules (Fig. 3; online supplemental table S4), as at the 90% target the criteria rule over all findings and the screening-restricted rule each called 60.4% of normal athletes abnormal, and the measured-intervals hybrid 58.2%, against 46.7% for the model's own output. The hybrid failed for a plain reason: the open delineator measured the athletes' median QRS duration at 111 to 112 ms, and on its own the measurement layer flagged 8 of the 28 endurance athletes and 47 of the 156 footballer recordings, so the measured half of the criteria was no more reliable than the model's named findings.

A readout trained on the athlete criteria does

The trained readouts behaved differently from the rules. At matched or higher sensitivity they called far fewer normal athletes abnormal along the whole operating curve (Fig. 3), but the pre-registered step-3 test of the neural readout failed on its sensitivity margin, because its thresholds had been set on folds that included its own training data, so its realised sensitivity fell below the target (0.840 against 0.878). The confirmatory step-3b test corrected this design flaw, and its hypothesis was met (Table 1). With thresholds set on a fold it had never seen, the readout called abnormal 17.6% (10.3 to 25.4) of normal athletes against 52.8% (43.4 to 62.4) for the model's own output, a difference of 35.2 percentage points (26.8 to 44.0), while its screening-red sensitivity was 87.5% against 89.4% on PTB-XL and 98.1% against 97.1% on Chapman/Ningbo. The result held for the two other training seeds (15.9% and 22.0%), it was largest in the endurance athletes (1 of 26, 3.8%, against 46.2%) and clear in the footballers (19.9% against 53.8%), and it carried over to non-athletes, with green-only false alarms falling from 41.4% to 15.7% in PTB-XL and from 20.3% to 10.0% in Chapman/Ningbo, although normal Chapman/Ningbo ECGs were flagged slightly more often (14.6% against 13.1%). At the 95% point the readout reached 94.9% sensitivity against 92.2% for the model's output while still calling fewer normal athletes abnormal (35.7% against 62.6%). The logistic readout lay between the two (26.4% at 90.0% sensitivity).

Confirmatory test (step 3b). Thresholds give 90% sensitivity for screening-relevant abnormal ECGs on PTB-XL fold 8, which no readout was trained on. Athlete rates pool 26 endurance athletes and 156 footballer recordings (95% athlete-level bootstrap intervals).

2–3(lr)4–6	Screening-red sensitivity		Normal athletes called abnormal			PTB-XL green-only
	PTB-XL folds 9–10	Chapman/Ningbo	pooled	endurance	footballers	called abnormal
Model's own output	0.894	0.971	0.528 [0.434, 0.624]	0.462	0.538	0.414
Neural readout, seed 0 (primary)	0.875	0.981	**0.176** [0.103, 0.254]	0.038	0.199	0.157
Neural readout, seed 1	0.878	0.975	0.159 [0.090, 0.235]	—	—	0.145
Neural readout, seed 2	0.880	0.986	0.220 [0.137, 0.305]	—	—	0.166
Logistic readout	0.900	0.951	0.264 [0.184, 0.351]	—	—	0.234

Figure 3.

Operating curves for the model's own output and the four fixes, with sensitivity for screening-relevant abnormal ECGs on PTB-XL test folds against the share of normal athletes called abnormal (step 3; points are the 80, 85, 90 and 95% targets). Large markers with 95% intervals: the confirmatory test (step 3b) with held-out thresholds; star, the neural readout; circle, the model's own output.

Discussion

Main findings

At a threshold that flags one in ten normal ECGs from the general population, a widely available ECG foundation model calls a third to a half of normal athletes abnormal and sometimes reports a heart attack, and this is not a thresholding accident, as the same pattern appears in non-athletes who carry the same findings and in two independent athlete groups from different sports and countries. Unlike our earlier conclusion that cheap fixes could not repair it, the confirmatory test showed that a small readout trained against the International Criteria on non-athletes removes about two-thirds of the false alarms in normal athletes with almost no loss of sensitivity for the findings screening is meant to catch, whereas rules applied over the model's named findings, or over intervals measured by an open delineator, made things worse.

Where the error comes from, and why the readout works

Taken together, the fixes locate the problem. The model's internal representation already holds the information needed to tell athletic adaptation from screening-relevant disease, since a readout trained on it with the right labels separates the two in non-athletes and carries that separation over to athletes it has never seen, which points to the output layer rather than the representation. The model's named findings are where the error sits, as a slow heart rate comes with an ST label, high voltage comes with hypertrophy and incomplete right bundle branch block comes with complete block, and this is faithful learning of the conventions in its training statements, much as PTB-XL itself files incomplete right bundle branch block under conduction disturbance and high voltage under hypertrophy. It can be inferred that our earlier readout failed because of what it was asked to detect rather than what it could see, as anchoring sensitivity on every non-green label, most of which in a hospital population are atrial fibrillation, ectopy and conduction disease in older patients, forced its threshold into the range of normal ECGs, whereas anchoring on the short list of screening targets did not.

These findings are in keeping with the history of athlete screening, in which the false-positive burden of general rules fell only when the criteria themselves were rewritten for athletes[1,2,4], and they echo the multicentre study of a general-purpose language model on athletes' ECGs[9], although that model mostly missed abnormalities while this one over-called normal ones. Both point the same way, that general medical artificial intelligence does not inherit the sports-cardiology conventions that took two decades to build.

What it means clinically

Green findings are the rule rather than the exception in screening, and sinus bradycardia is the most common athletic ECG finding, which is exactly where the model's error is concentrated, below 50 beats per minute. At the confirmatory operating point, a pathway that put the model's own output first would refer about 528 of every 1,000 normal athletes for further tests, the very cascade the athlete criteria were written to prevent, while the criteria-trained readout would refer about 176, with wide intervals because the athlete groups are small. Neither figure is yet acceptable for unsupervised screening, and conventional machine reading already over-calls athletes, so a system presented as more accurate than the conventional algorithm risks being trusted where the old one was not. The practical lesson is cheaper than we expected, however: foundation models do not need to be retrained to stop treating athletic adaptation as disease, but their readouts need to be labelled against the athlete criteria and validated on athletes before they are used.

Comparison of the fixes

The measured-intervals hybrid was the design closest to how a sports cardiologist reads an ECG, measuring what can be measured and asking the model only about morphology, and its failure is itself a finding, because it shows that an open delineator is not accurate enough to carry the measured half of the criteria, since its QRS and corrected-QT rules fired on athletes a cardiologist had read as normal. A

device-grade measurement engine might change this, but none was available to us, and the GE 12SL readings available for the endurance athletes were themselves borderline or abnormal in half of them.

Limitations

This study has several limitations. The athlete groups are small (26 athletes and 52 players), so the intervals are wide, and the athletes were recorded on GE devices whereas the readout and thresholds came from PTB-XL, so device or site differences could contribute to their rates, although the confirmatory comparison is within the same data for both methods. The readout was trained and thresholded on hospital ECGs, and its sensitivity was measured on screening-relevant abnormal ECGs from older patients, with Q-wave infarction as a proxy for pathological Q waves, so its sensitivity for hypertrophic cardiomyopathy, arrhythmogenic cardiomyopathy and Brugada patterns in young athletes, which these datasets do not label, is untested. The confirmatory test followed a pre-registered step whose sensitivity test failed for a design reason, and although the confirmatory test was written before its data were scored, the choice of readout came from the earlier step.

We tested one foundation model, as ECG-FM[14] needs a separate software toolchain and HuBERT-ECG[15] was excluded because its training data include both non-athlete test datasets. The Black-athlete repolarisation variant was not tested, because no open dataset labels it, footballers with anterior T-wave inversion were excluded rather than guessed, the African-ancestry group ($n = 5$) is descriptive only, and sex was not analysed. Finally, the PF12RED findings are marks per player rather than a structured reading against the International Criteria, the PF12RED files carry no device interpretation, and the sinus-bradycardia difference between PTB-XL and Chapman/Ningbo probably reflects label definitions and heart-rate distributions, which we did not resolve.

Declarations

Contributors CE conceived and designed the study, performed the analysis, interpreted the results and wrote the manuscript. CE is the guarantor and accepts full responsibility for the work and the conduct of the study, had access to the data, and controlled the decision to publish.

Competing interests

None declared.

Patient and public involvement

Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

Ethics approval

This study analysed only publicly available, fully de-identified datasets released under open licences (PTB-XL, Chapman/Ningbo, Norwegian Endurance Athlete ECG Database, PF12RED). It involved no contact with participants and no access to identifiable information, so it does not constitute human-subjects research and no ethics approval or exemption was required.

Equity, diversity and inclusion statement

The athlete data come from elite endurance athletes in Norway and professional footballers in Spain, and only five footballers were of African ancestry. No open dataset labels the repolarisation variant seen in Black athletes, so we excluded affected footballers rather than guess, and our findings may not generalise to Black athletes, to athletes outside Europe, or across sexes, which we did not analyse. The author is a physician based in Nigeria, and the work is motivated by the use of free, open models in settings without specialist sports cardiology.

Use of generative AI

Claude (Anthropic) assisted with analysis code and drafting; the author reviewed all content and takes full responsibility for it.

Funding

The author has not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Data availability statement

Data are available in a public, open access repository. PTB-XL and Chapman/Ningbo are available from PhysioNet; the Norwegian Endurance Athlete ECG Database (CC BY 4.0) and PF12RED (CC0) are openly licensed; ECGFounder is MIT-licensed. Pre-registrations for all four steps, code and per-ECG outputs are available at https://github.com/Harry-Potter20/paper-code/tree/main/green_means_go.

References

1. Corrado D, Pelliccia A, Heidbuchel H et al (2010) Recommendations for interpretation of 12-lead electrocardiogram in the athlete. Eur Heart J 31(2):243–259. 10.1093/eurheartj/ehp473
2. Drezner JA, Ackerman MJ, Anderson J et al (2013) Electrocardiographic interpretation in athletes: the ‘Seattle criteria’. Br J Sports Med 47(3):122–124. 10.1136/bjsports-2012-092067

3. Sheikh N, Papadakis M, Ghani S et al (2014) Comparison of electrocardiographic criteria for the detection of cardiac abnormalities in elite black and white athletes. Circulation 129(16):1637–1649. 10.1161/CIRCULATIONAHA.113.006179
4. Sharma S, Drezner JA, Baggish A et al (2017) International recommendations for electrocardiographic interpretation in athletes. J Am Coll Cardiol 69(8):1057–1075. 10.1016/j.jacc.2017.01.015
5. Vardar U, Chung EH (2026) International criteria for ECG interpretation in athletes. American College of Cardiology, Expert Analysis; 29 October 2026
6. Li J, Aguirre AD, Moura Junior V et al (2025) An electrocardiogram foundation model built on over 10 million recordings. NEJM AI 2(7). 10.1056/AIoa2401033
7. Singstad BJ et al (2022) Norwegian Endurance Athlete ECG Database. IEEE Open J Eng Med Biol 3:162–166. 10.1109/OJEMB.2022.3214719
8. Munoz-Macho AA, Domínguez-Morales MJ, Sevillano-Ramos JL (2024) An innovative 12-lead resting electrocardiogram dataset in professional football. Data Brief 54:110444. 10.1016/j.dib.2024.110444
9. Palermi S, Vecchiato M, Iacovone T et al (2026) ChatGPT's limitations in athlete ECG interpretation: evidence from a multicenter diagnostic study. J Cardiovasc Dev Dis 13(5):191. 10.3390/jcdd13050191
10. Wagner P, Strodthoff N, Bousseljot RD et al (2020) PTB-XL, a large publicly available electrocardiography dataset. Sci Data 7:154. 10.1038/s41597-020-0495-6
11. Zheng J, Zhang J, Danioko S et al (2020) A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. Sci Data 7:48. 10.1038/s41597-020-0386-x
12. Zheng J, Guo H, Chu H (2022) A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0.0). 10.13026/wgex-er52. PhysioNet
13. Collins GS, Moons KGM, Dhiman P et al (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. 10.1136/bmj-2023-078378
14. McKeen K, Masood S, Toma A et al (2025) ECG-FM: an open electrocardiogram foundation model. JAMIA Open 8:ooaf122. 10.1093/jamiaopen/ooaf122
15. Coppola E, Savardi M, Massussi M et al (2024) HuBERT-ECG as a self-supervised foundation model for broad and scalable cardiac applications. medRxiv. 10.1101/2024.11.14.24317328

Figures

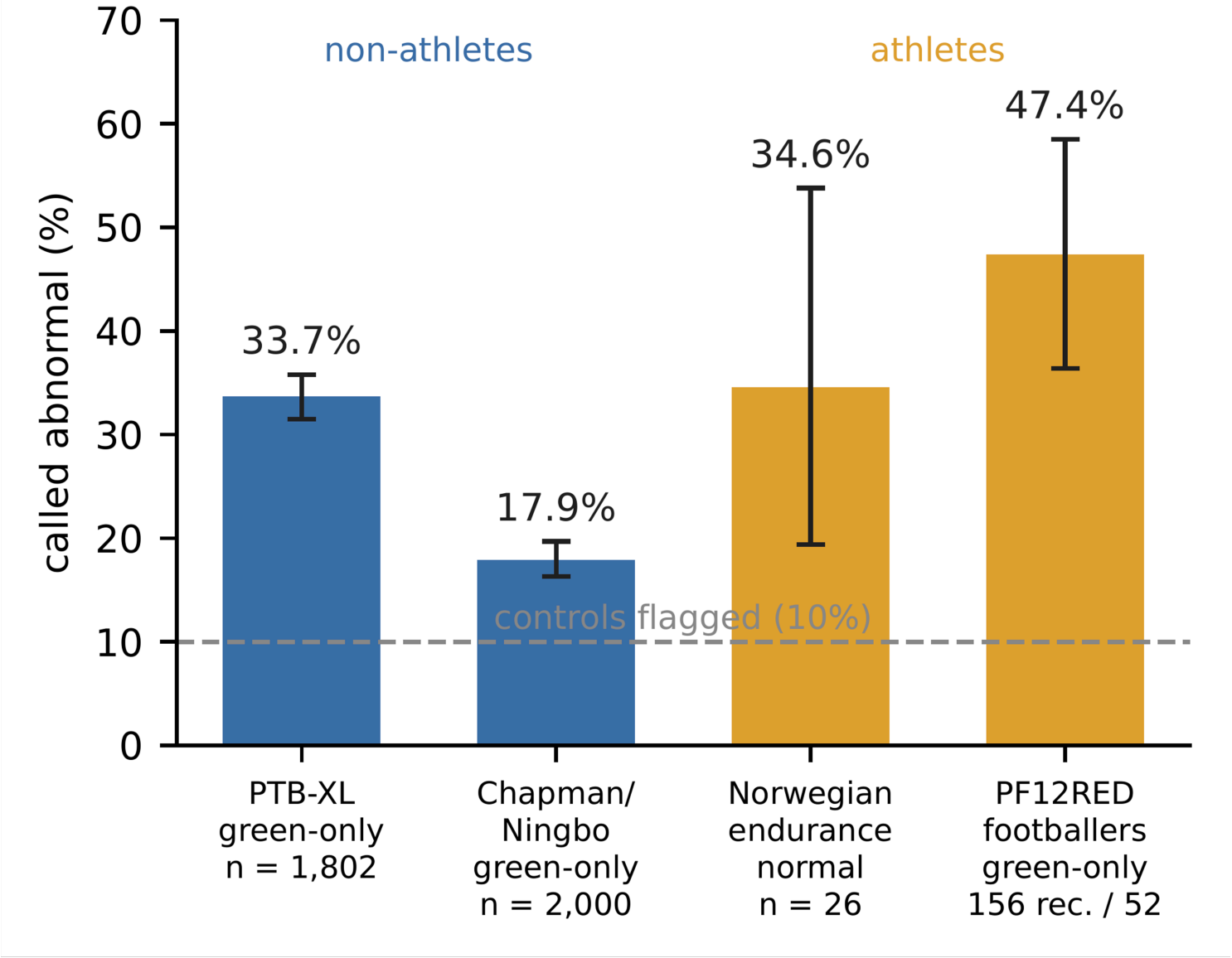


Figure 1

Share of ECGs called abnormal at an operating point that flags 10% of normal ECGs (dashed line). Blue: non-athletes with green findings only. Orange: normal athletes. Bars are 95% intervals (Wilson; footballers by player-level bootstrap).

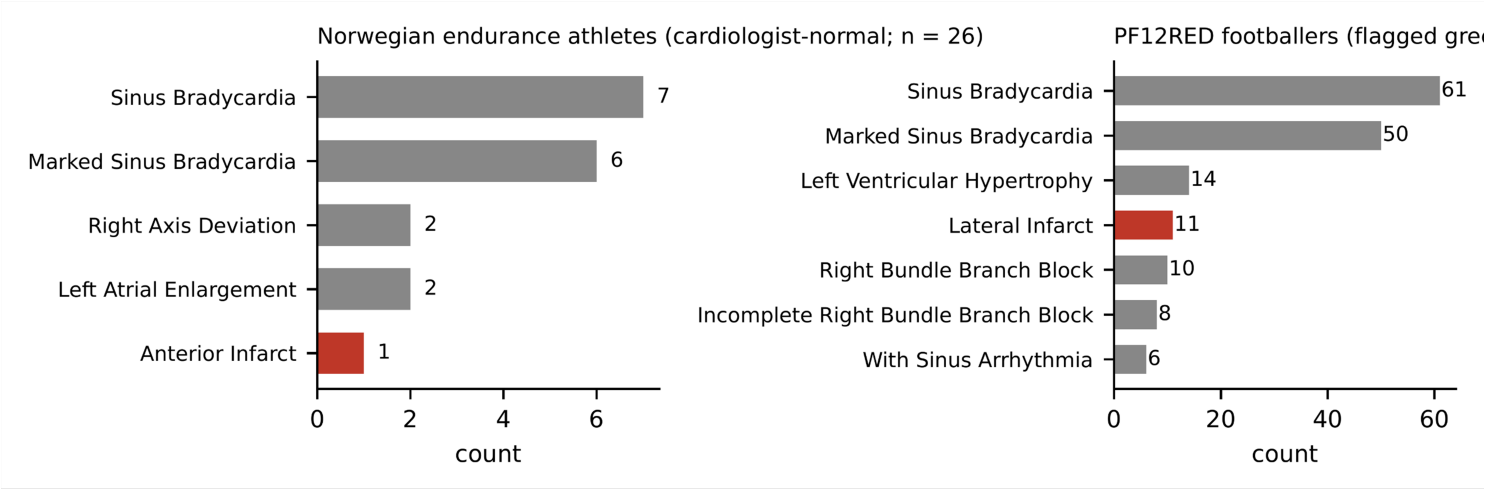


Figure 2

Abnormal findings the model assigned to normal athletes. Infarction labels are shown in red.

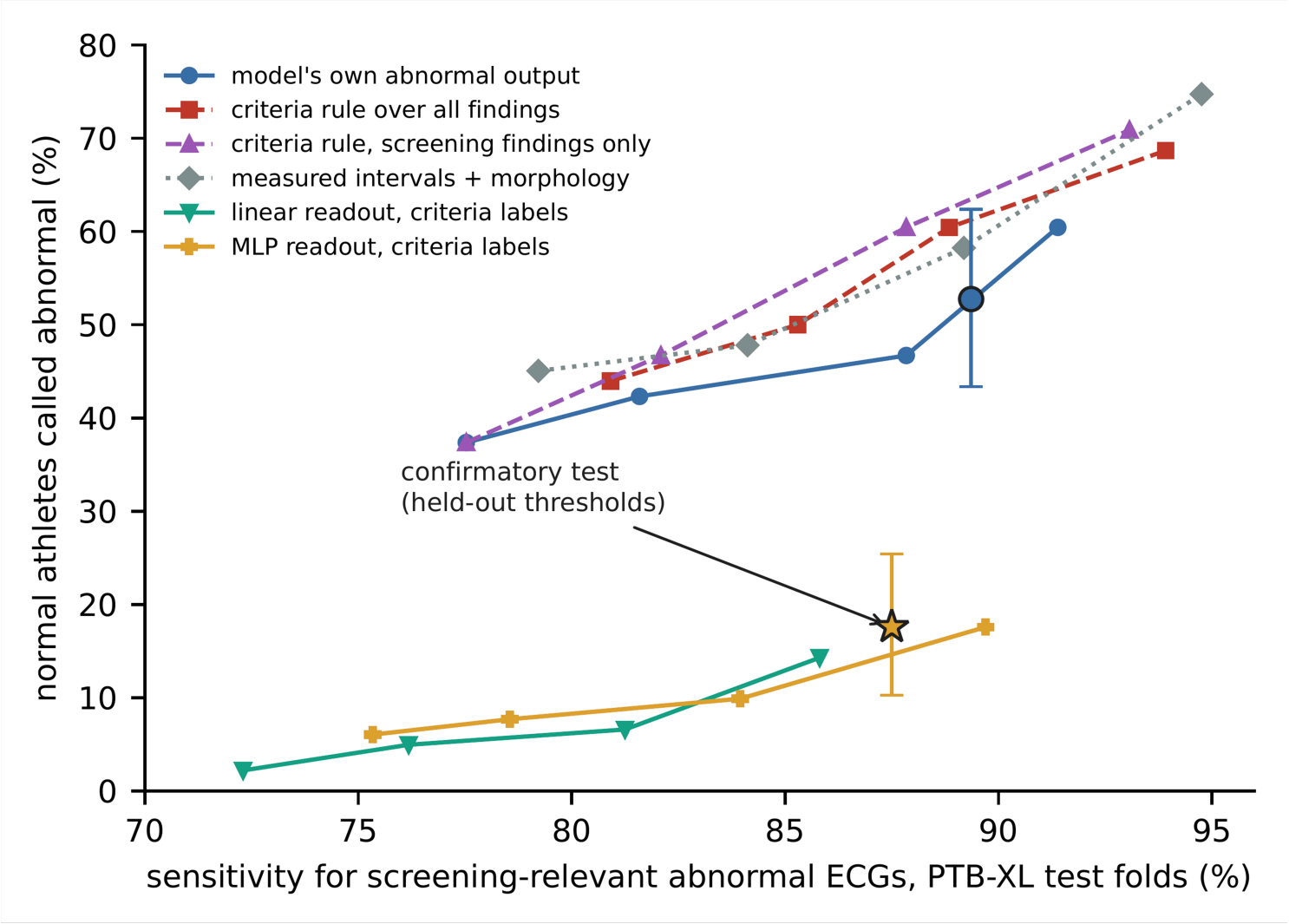


Figure 3

Operating curves for the model’s own output and the four fixes, with sensitivity for screening-relevant abnormal ECGs on PTB-XL test folds against the share of normal athletes called abnormal (step 3; points are the 80, 85, 90 and 95% targets). Large markers with 95% intervals: the confirmatory test (step 3b) with held-out thresholds; star, the neural readout; circle, the model’s own output.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarytables.tex
- TRIPODAIchecklist.tex