ECG2ACT: FROM ABNORMALITY RECOGNITION TO CLINICAL ACTION CONCERN PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electrocardiogram (ECG) analysis informs not only the diagnostic evaluation of cardiac abnormalities but also care planning, which often must proceed before a definitive diagnosis is established. Yet existing ECG deep learning methods mainly focus on abnormality recognition, without explicitly identifying the diagnostic and care-process concerns that guide subsequent clinical assessment and management. We therefore introduce ***clinical action concern prediction***, which jointly identifies problem-oriented diagnostic concerns and care-process concerns from a single ECG to inform subsequent clinical assessment and management. This task presents **two challenges**: the two target families share the same ECG input yet may rely on *different waveform features*, while the concern-level supervision indicates *which concerns are present* without specifying *which ECG characteristics support them*. To address these challenges, we develop **ECG2Act**, which learns a shared ECG representation and uses concern-specific action queries to retrieve target-relevant evidence, while utilizing measurement regression and evidence classification supervision to anchor this representation to interpretable ECG characteristics. On patient-disjoint evaluation of **179,100 ECGs from 82,486 subjects**, ECG2Act achieves diagnostic and care-process AUROCs of **0.938 and 0.713**, respectively. Crucially, care-process concerns remain distinguishable among ECGs sharing the same diagnostic profile, demonstrating that ECG-based care assessment captures information not represented by diagnostic recognition alone. These results support extending current ECG analysis beyond abnormality recognition to jointly assess diagnostic and care-process action concerns.

1 INTRODUCTION

The electrocardiogram (ECG) informs both the assessment of cardiac problems and the planning of subsequent care (Rao et al., 2025; Joglar et al., 2024; Heidenreich et al., 2022). Deep learning has advanced mainly along the first dimension, from morphology-based heartbeat classification (de Chazal et al., 2004) to learned arrhythmia recognition (Mousavi et al., 2018; Hannun et al., 2019) and multi-abnormality diagnosis from 12-lead ECGs (Ribeiro et al., 2020). More recent ECG grounding systems, including GEM and ECG-R1, further link abnormality statements to waveform evidence (Lan et al., 2025; Jin et al., 2026). These advances improve recognition and explanation, but abnormality identification alone does not capture the full set of concerns arising during clinical assessment. Related findings may point to the same cardiac problem yet imply different needs for subsequent care. Thus, current ECG analysis does not explicitly represent both the cardiac problems under assessment and the care-process concerns associated with the recording.

This gap matters because care planning often begins before diagnostic uncertainty is resolved (Gulati et al., 2021; Rao et al., 2025). A single ECG may remain diagnostically inconclusive while still raising concerns about *serial ECG reassessment* for evolving changes, *cardiac monitoring* for continued surveillance, *admission or observation* for further evaluation, or *high-acuity care* when closer management is warranted (Rao et al., 2025; Sandau et al., 2017; Gulati et al., 2021). These concerns reflect distinct aspects of subsequent assessment and management, not successive degrees of ECG abnormality. ECG analysis should therefore identify both the cardiac problems warranting attention and the care-process concerns relevant to what follows. This motivates a broader question: ***how can a single ECG jointly inform diagnostic and care-process action concerns?***

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

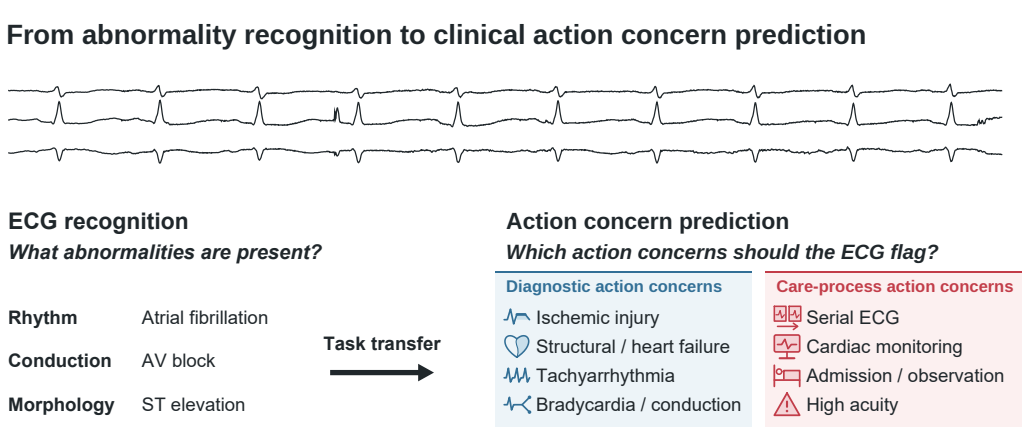


Figure 1: **From abnormality recognition to clinical action concern prediction.** Diagnostic concerns group related findings by cardiac problem, while care-process concerns address reassessment, monitoring, and the setting and intensity of subsequent care. Both are assessed from a single ECG.

Answering this broader clinical question from a single ECG introduces two learning challenges. ***The first is concern-specific feature learning.*** Diagnostic and care-process concerns share physiology but address different clinical questions: features that distinguish cardiac problems may not be equally informative about subsequent care. Joint learning must exploit shared ECG structure while allowing each concern to draw on the waveform features relevant to its assessment (Caruana, 1997; Liu et al., 2021). ***The second is physiological grounding.*** Concern-level labels indicate which concerns are present, but do not specify the waveform characteristics that support them. Rates, intervals, and recognizable waveform patterns inform ECG interpretation (Kligfield et al., 2007; Mason et al., 2007), yet concern labels alone do not explicitly supervise these properties. Measurements and findings offer explicit physiological supervision but capture only selected aspects of the waveform. The challenge is to use these targets to guide learning while retaining access to other relevant features.

We formulate ***clinical action concern prediction*** as the joint assessment of problem-oriented diagnostic and care-process concerns from a single ECG (Figure 1), and develop **ECG2Act** to address these learning challenges. A shared *convolutional-Transformer* encoder produces joint-lead waveform tokens, from which eight learnable action queries retrieve concern-specific features. To provide physiological grounding, measurement regression supervises the pooled encoder representation, while evidence classification supervises pooled query representations using rhythm, conduction, and morphology findings. These objectives guide feature learning during training; the concern heads predict each concern directly from its query representation.

We construct a cohort of **179,100 ECGs from 82,486 subjects** and evaluate ECG2Act with patient-disjoint splits, obtaining macro-AUROCs of **0.938 and 0.713** for diagnostic and care-process concerns, respectively. Beyond family-level performance, we test whether care predictions merely reflect the four diagnostic labels. Care discrimination persists among recordings with identical diagnostic labels, indicating that these labels do not capture all care-related information learned from ECG waveforms. Auxiliary-supervision ablations validate the contribution of physiological guidance, while longitudinal analyses and waveform controls assess how predictions respond to changes across recordings and to altered ECG structure. Together, these results support extending ECG analysis beyond abnormality recognition to jointly assess both diagnostic and care-process concerns.

Our contributions cover task formulation, the learning framework, and their empirical evaluation:

- We formulate *clinical action concern prediction*, jointly assessing problem-oriented diagnostic and care-process concerns from ECGs, and construct a large-scale cohort for evaluation.
- We develop **ECG2Act**, coupling concern-specific action queries with physiological auxiliary supervision to learn shared waveform features for joint diagnostic and care-process prediction.
- ECG2Act achieves diagnostic and care-process macro-AUROCs of **0.938 and 0.713**, respectively, under patient-disjoint evaluation on **179,100 ECGs from 82,486 subjects**.

2 Related Work

ECG representations and physiological grounding. Supervised models address arrhythmia recognition and multi-abnormality diagnosis directly from ECG waveforms (Hannun et al., 2019; Ribeiro et al., 2020). To reduce reliance on task-specific labels, CLOCS learns transferable features through contrasts across leads, time, and patients (Kiyasseh et al., 2021), while ECG-FM combines masked and contrastive learning for transfer across ECG tasks (McKeen et al., 2025). More recently, GEM links diagnostic statements to measurable waveform evidence (Lan et al., 2025), and ECG-R1 uses protocol-guided instruction data and diagnostic-evidence rewards to improve interpretation (Jin et al., 2026). These advances improve diagnostic representations and explanations, but do not explicitly assess the care-process concerns relevant to subsequent clinical assessment and management. ECG2Act extends this physiological emphasis to diagnostic and care-process assessment, using auxiliary measurements and ECG findings supervision to guide shared representations.

Care prediction and joint clinical assessment. Triage models predict hospitalization and critical care from routine clinical variables (Raita et al., 2019), while multimodal ICU models combine ECG waveforms with vital signs and laboratory results to predict deterioration (López Alcaraz et al., 2026). ECG-based mortality alerts also connect waveform prediction to clinical intervention (Lin et al., 2024). Closer to our setting, Altintepe & Ozyoruk (2025) combine serial ECGs with clinical data for admission prediction, include ECG-only baselines, and pretrain on measurements and report labels. ECG-based care prediction and physiological pretraining therefore have clear precedents. ECG2Act instead studies the joint assessment of problem-oriented diagnostic and care-process concerns using ECGs as the only input. We further test care discrimination among recordings sharing the same diagnostic labels to assess care-related information beyond these selected categories.

3 Method

3.1 Clinical Action Concern Prediction

For a 10-s, 12-lead ECG sampled at 500 Hz, $\mathbf{X} \in \mathbb{R}^{12\times 5000}$, let $\mathbf{y} = [\mathbf{y}^{\text{diag}}; \mathbf{y}^{\text{care}}] \in \{0,1\}^K$ denote $K = 8$ nonexclusive labels covering four problem-oriented diagnostic concerns and four observed care-process concerns (Table 1). ECG2Act predicts all eight concern scores,

$$\mathbf{p} = f_\theta(\mathbf{X}) = [\mathbf{p}^{\text{diag}}; \mathbf{p}^{\text{care}}] \in [0,1]^K.$$

A shared encoder and concern-specific queries implement this mapping (Figure 2); physiological targets (measurements and findings) guide feature learning without being inputs to concern prediction.

3.2 Joint-Lead Waveform Encoding

We standardize each lead using its temporal mean μ_c and standard deviation σ_c within the recording: $\widetilde{X}_{ct} = (X_{ct} - \mu_c)/(\sigma_c + 10^{-6})$. A four-stage convolutional stem mixes lead channels and downsamples time to produce $N = 157$ joint-lead tokens of dimension $d = 512$. Learnable position embedding $\mathbf{P}_{1:N} \in \mathbb{R}^{N\times d}$ are added before six pre-LayerNorm Transformer blocks:

$$\mathbf{U} = \text{Stem}(\widetilde{\mathbf{X}})^\top, \qquad \mathbf{H} = \text{LN}(\text{Enc}_6(\mathbf{U} + \mathbf{P}_{1:N})) \in \mathbb{R}^{N\times d}. \tag{1}$$

Transformer encoders model temporal context (Vaswani et al., 2017), while the full sequence $\mathbf{H}$ is retained for concern-specific aggregation. Encoder settings are detailed in Appendix B (Table A1).

3.3 Concern-Specific Action Queries

Each row of $\mathbf{Q} \in \mathbb{R}^{K\times d}$ is a learned query for one concern. Multihead cross-attention uses $\mathbf{H}$ as keys and values, followed by query residuals, LayerNorm, and affine heads:

$$\mathbf{A} = \text{LN}(\mathbf{Q} + \text{MHA}(\mathbf{Q}, \mathbf{H}, \mathbf{H})), \qquad s_k = \mathbf{w}_k^\top \mathbf{a}_k + b_k, \quad p_k = \sigma(s_k). \tag{2}$$

Here $\mathbf{a}_k^\top$ is row k of $\mathbf{A} \in \mathbb{R}^{K\times d}$, $\mathbf{w}_k \in \mathbb{R}^d$, $b_k \in \mathbb{R}$, and σ is sigmoid. Attention is normalized over tokens rather than concerns (Liu et al., 2021). Each query can therefore retrieve shared ECG waveform evidence for its own concern label, while sigmoid heads allow nonexclusive predictions.

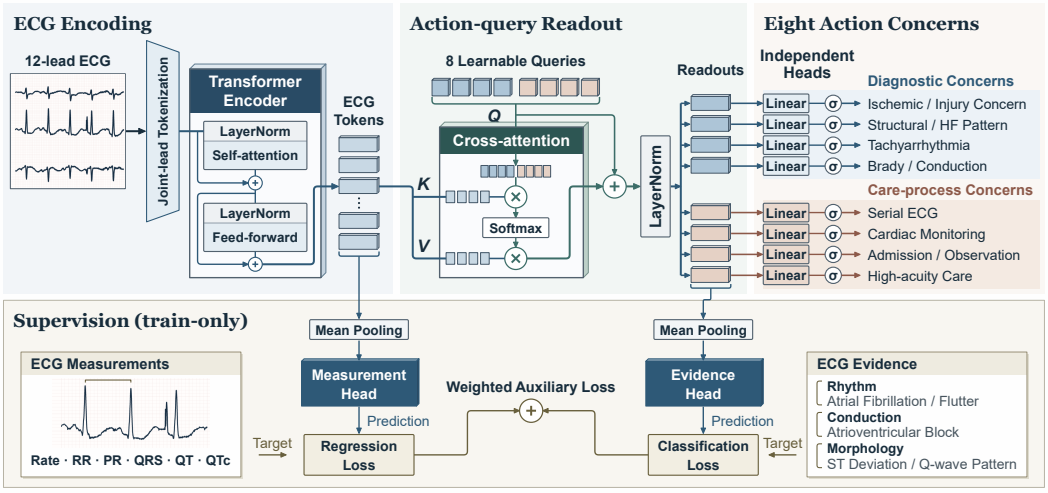


Figure 2: ECG2Act couples shared waveform encoding to concern-specific readout. Eight queries produce diagnostic and care scores; measurement and finding objectives supervise pooled encoder and query representations, respectively. Auxiliary branches are used only during training.

3.4 Physiological Supervision

Action labels supervise the final concerns; auxiliary targets explicitly supervise ECG properties. Measurement regression uses pooled encoder tokens, and finding classification averages all $K = 8$ normalized query readouts to predict ECG findings:

$$\widehat{\mathbf{m}} = g_M\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{h}_n\right) \in \mathbb{R}^9, \qquad \mathbf{e} = g_E\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{a}_k\right) \in \mathbb{R}^{14}. \tag{3}$$

Here $\mathbf{h}_n^\top$ is row n of $\mathbf{H}$; $\widehat{\mathbf{m}}$ contains standardized measurement predictions and $\mathbf{e}$ contains finding logits. Each head is a two-layer MLP with a 512-unit GELU hidden layer. Measurements comprise *heart rate, RR, PR, QRS, QT, QTc, and the P, QRS, and T axes*. The fourteen findings span rhythm, conduction, and morphology, including *atrial fibrillation/flutter, atrioventricular block, ST deviation, and Q-wave patterns*. Measurements supervise the shared encoder; findings additionally supervise the action-query representations. More information is detailed in Appendix A.

3.5 Joint Learning Objective

For a minibatch of B recordings, s_{ik} and y_{ik} denote the logit and label for concern k in recording i. Training-split positive and negative counts n_k^+, n_k^- define capped positive-class weights:

$$\alpha_k = \mathrm{clip}\left(\frac{n_k^-}{\max(n_k^+, 1)}, 0.1, 2\right), \qquad \mathcal{L}_{\mathrm{act}} = \frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K}\ell_{\mathrm{BCE}}(s_{ik}, y_{ik}; \alpha_k). \tag{4}$$

Here $\ell_{\mathrm{BCE}}(s, y; \alpha) = -\alpha y \log\sigma(s) - (1-y)\log(1-\sigma(s))$: only the positive term is reweighted, and all BK labels contribute to the mean. Measurements m_{ir}, $r \in \{1, \ldots, 9\}$, use training-split standardization and binary availability masks M_{ir}. Findings z_{ij}, $j \in \{1, \ldots, 14\}$, have fixed weights $W_{ij} \geq 0$, with unavailable entries assigned zero rather than a negative label. The auxiliary losses are

$$\mathcal{L}_{\mathrm{meas}} = \frac{\sum_{i,r} M_{ir}\rho_1(\widehat{m}_{ir} - m_{ir})}{\max(\sum_{i,r} M_{ir}, 1)}, \qquad \mathcal{L}_{\mathrm{evid}} = \frac{\sum_{i,j} W_{ij}\ell_{\mathrm{BCE}}(e_{ij}, z_{ij}; 1)}{\max(\sum_{i,j} W_{ij}, 1)}. \tag{5}$$

Here ρ_1 is unit-transition Smooth L1. Both reductions pool observed entries across the minibatch; finding BCE uses unit positive-class weight. Physiological and action supervision are combined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{act}} + 0.20\,\mathcal{L}_{\mathrm{meas}} + 0.05\,\mathcal{L}_{\mathrm{evid}}. \tag{6}$$

We optimize with AdamW (Loshchilov & Hutter, 2019) and maintain a parameter exponential moving average with decay 0.995. Further implementation details appear in Appendix B.

4 EXPERIMENTS

Table 1: Eight nonexclusive action concerns and observed positive rates. Diagnostic targets group cardiac problems; care targets describe recorded events or settings. HF abbreviates *heart failure*.

Action concern	Positive rate	Target meaning
Diagnostic action concerns		
Ischemic injury	21.5%	ST/T-wave abnormalities or infarct patterns
Structural / HF	37.4%	Structural ECG patterns or encounter-linked HF
Tachyarrhythmia	21.6%	Atrial fibrillation/flutter or tachyarrhythmias
Brady / conduction	22.8%	Slow rhythm, conduction block, or widened QRS
Care-process action concerns		
Serial ECG	31.1%	Repeat ECG assessment after the index recording
Cardiac monitoring	34.9%	Transfer to a unit providing cardiac monitoring
Admit / observe	63.9%	Hospital admission or an observation stay
High acuity	11.4%	Admission or transfer to intensive care

Table 2: Recording-level performance on the evaluation partition. Macro rows weight concerns equally; F1 uses validation-selected thresholds (Table A2). Positive rates are given in Table 1.

Action concern	AUROC	AUPRC	F1
Diagnostic action concerns			
Ischemic injury	0.922	0.801	0.707
Structural / HF	0.862	0.796	0.725
Tachyarrhythmia	0.990	0.971	0.919
Brady / conduction	0.977	0.936	0.861
Diagnostic macro	**0.938**	**0.876**	**0.803**
Care-process action concerns			
Serial ECG	0.677	0.510	0.516
Cardiac monitoring	0.656	0.517	0.538
Admit / observe	0.719	0.806	0.791
High acuity	0.801	0.383	0.412
Care macro	**0.713**	**0.554**	**0.564**
Overall macro	**0.825**	**0.715**	**0.684**

4.1 EVALUATION DESIGN

We evaluate ECG2Act through concern discrimination, auxiliary-supervision ablation, and analyses of the information retained in its predictions. Tables 1 and 2 define the eight targets and report concern-wise performance; Table 3 compares measurement and finding supervision under a common architecture. We then examine care discrimination within diagnostic groups, changes across successive ECGs, and responses to controlled perturbations of waveform timing and cardiac morphology.

Data and supervision. We use MEETI (Zhang et al., 2026), derived from MIMIC-IV-ECG (Gow et al., 2023), for ECG signals, and matched ECG measurements and findings for auxiliary supervision. Emergency visits and hospital records supply care-process labels (Johnson et al., 2023; 2024). The cohort has 179,100 ECGs from 82,486 subjects, with patient-disjoint splits of 143,885 training, 17,628 validation, and 17,587 evaluation recordings. Appendix A specifies each source, label construction, and supervision masks. The ECG waveforms are the only input during training and inference.

Evaluation protocol. Table 2 reports recording-level AUROC, AUPRC, and F1, with equal-weight macro averages for each family and all eight concerns. Separate family scores distinguish diagnostic performance from care prediction; target prevalence in Table 1 contextualizes precision–recall results. F1 thresholds are selected on validation and fixed for evaluation (Table A2). Subjects never cross splits. The evaluation partition was used to compare and select model configurations; the reported scores summarize patient-disjoint performance of the selected model under this evaluation protocol.

4.2 Discrimination and Auxiliary Supervision

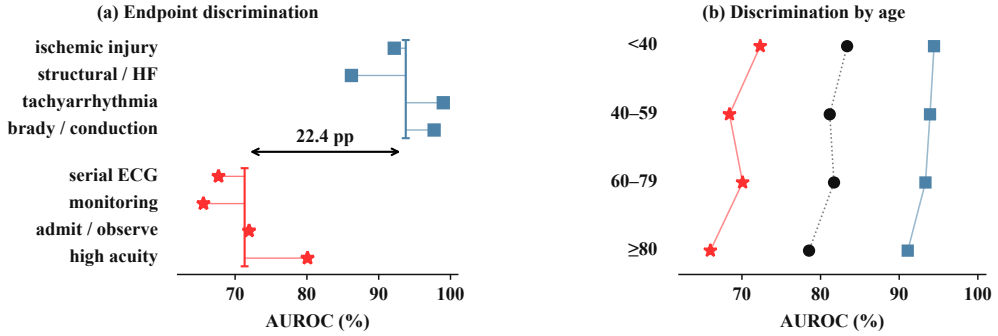


Figure 3: AUROC (%) across concerns and age groups. (a) Concern-level estimates with family means marked by vertical lines. (b) Age-stratified macro-AUROC. **Blue** squares, **Black** circles, and **Red** stars indicate diagnostic, overall, and care results, respectively.

Table 3: **Auxiliary-supervision ablation.**

Auxiliary supervision		AUROC		Macro average		
Measurements	Findings	**Diagnostic**	**Care**	AUROC	AUPRC	F1
○	○	0.9225	0.7018	0.8122	0.7015	0.6696
○	●	0.9279	0.7024	0.8152	0.7020	0.6765
●	○	0.9325	0.7077	0.8201	0.7084	0.6791
●	●	**0.9377**	**0.7133**	**0.8255**	**0.7151**	**0.6836**

A single ECG supports diagnostic and care-process assessment. We evaluate all eight concerns to test whether ECG2Act can learn the broader task from one ECG recording (Table 2). Diagnostic and care-process macro-AUROCs reach **0.938** and **0.713**, with macro-F1 scores of 0.803 and 0.564, respectively. Care-process AUROCs range from 0.656 for monitoring to 0.801 for high-acuity care, showing predictive value across all four care targets. The strongest care result links waveform information to the intensity of subsequent management. These findings support extending ECG analysis from abnormality recognition to joint assessment of problem-oriented diagnostic concerns and care-process concerns relevant to subsequent assessment and management.

The diagnostic–care gap persists across concerns and ages. We next examine whether the family-level difference reflects isolated targets or a consistent pattern across patient groups (Figure 3). Panel **(a)** shows that every diagnostic concern exceeds every care-process concern, producing a 22.4-point gap between family means. Panel **(b)** preserves this ordering within all four age groups, even as performance declines in patients aged 80 or above. Care discrimination remains above chance in every stratum, showing that care-related discrimination is not confined to one patient subgroup. The persistent gap nevertheless identifies care-process prediction as the harder component of joint assessment. The contrast is consistent with our feature-learning motivation: shared physiology need not make waveform characteristics equally informative for every concern.

Measurements and findings provide complementary supervision. To test physiological grounding, we vary the two auxiliary objectives while keeping the architecture and action-concern supervision unchanged (Table 3). Combining them raises diagnostic and care-process macro-AUROCs by **1.52 and 1.15 percentage points** over the action-only baseline. Measurement regression is stronger alone, but adding evidence classification still improves diagnostic and care-process AUROCs by a further 0.52 and 0.56 points. Notably, evidence classification alone improves care AUROC by only 0.06 points, suggesting greater value when measurements are already supervised. The combined setting also achieves the highest overall macro-AUPRC and macro-F1. This pattern suggests complementary benefits from quantitative ECG measurements and recognizable findings, rather than interchangeable supervision signals. The gains in both families support physiological supervision for learning features relevant to care-process prediction as well as diagnostic assessment.

4.3 Care Information Beyond Diagnostic Labels

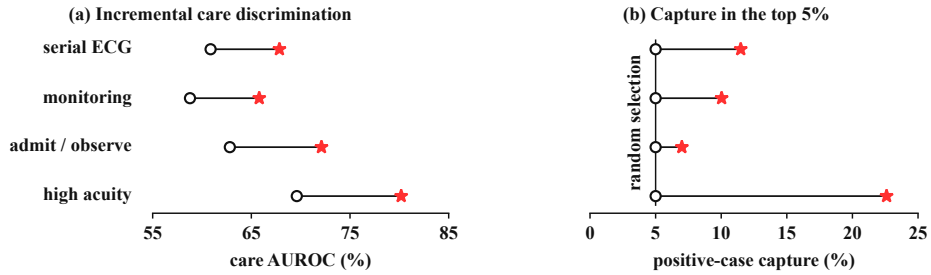


Figure 4: Care information and event concentration. (a) Diagnostic-label probes without (**Black** circles) and with ECG scores (**Red** stars). (b) Capture among all positive cases when each care score independently selects its top 5% (stars); circles denote random selection.

Table 4: Care AUROC (%) within identical diagnostic profiles. I: ischemic injury; S: structural/HF; T: tachyarrhythmia; B: brady/conduction. Unlisted labels are negative. The final row weights within-profile AUROCs by ECG count.

Diagnostic profile	ECGs	Serial ECG	Monitoring	Admit/obs.	High acuity
None positive	6,024	64.6	61.2	66.9	76.7
I	991	63.3	62.5	64.5	77.4
S	2,406	64.6	65.1	72.8	75.4
T	1,432	65.3	65.4	68.3	76.5
B	1,678	59.3	61.8	65.9	71.0
I + S	1,005	66.9	66.5	70.3	75.0
I + T	334	69.3	61.4	69.9	74.1
S + T	903	67.2	65.9	75.2	75.9
I + B	308	62.0	56.2	61.9	79.0
S + B	916	61.8	66.1	73.3	75.4
I + S + T	477	69.4	65.6	65.3	71.1
I + S + B	468	65.4	63.5	65.7	71.1
S + T + B	261	65.6	58.9	64.8	69.6
Sample-weighted mean	17,203	**64.4**	**63.1**	**68.5**	**75.3**

Care prediction adds information beyond diagnostic labels. To test whether care prediction extends beyond diagnostic associations, we compare validation-fitted probes using the four diagnostic labels alone or augmented with ECG-derived scores (Figure 4a). Augmentation improves all four care targets, raising macro-AUROC by **8.46 percentage points**. ECG-derived predictions thus retain care-relevant information beyond the diagnostic summary already available to the probe.

Care concerns remain distinguishable within the same diagnostic profile. We next compare recordings with identical four-label profiles, so that care discrimination cannot be attributed to differences in those labels (Table 4). Across 13 profiles comprising 17,203 ECGs, sample-weighted within-profile AUROCs range from 0.631 for monitoring to **0.753** for high-acuity care. High-acuity discrimination also reaches 0.767 among 6,024 ECGs negative for all four diagnostic concerns. A predictor restricted to the four labels assigns identical scores within each group. ECG2Act's residual discrimination therefore supports predicting care concerns directly from waveform representations rather than using the selected diagnostic labels as their sole basis.

Care scores concentrate relevant events within a small review subset. We next ask whether the additional information helps prioritize recordings when only a small fraction can be reviewed (Figure 4b). Selecting each care score's top 5% captures more events than random selection for all four targets. For high-acuity care, it captures more than 20% of positive cases, over four times the random-selection rate. Enrichment is smaller for admission/observation, indicating that event concentration differs across care processes. These rankings identify subsets enriched for specific care events, supporting concern-specific priorities for further review alongside diagnostic assessment.

4.4 Patient, Longitudinal, and Waveform Analyses

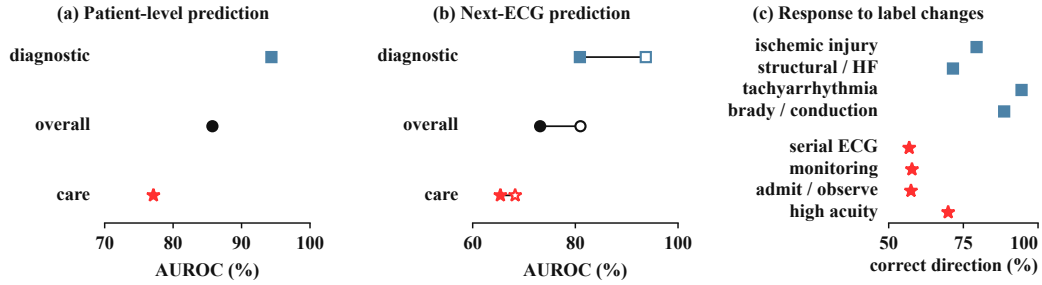


Figure 5: Patient and longitudinal analyses. (a) Per-concern maximum scores versus any-positive patient labels. (b) The same current scores evaluated against current (open) or next-ECG (filled) labels. (c) Direction agreement on label-changing pairs. **Blue**: diagnostic; **Black**: overall; **Red**: care.

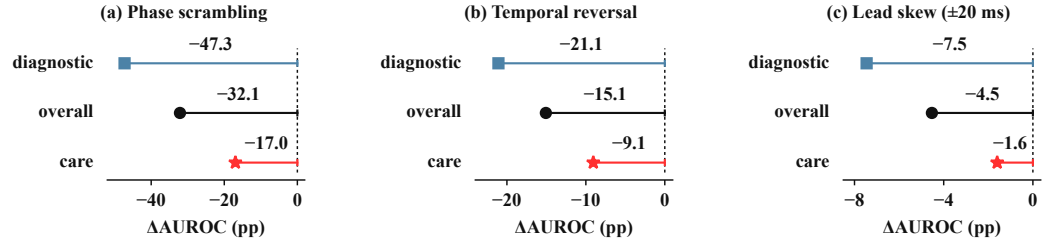


Figure 6: Perturbation-induced AUROC changes in percentage points, computed as perturbed minus original. Negative values indicate lower discrimination; horizontal scales differ across panels. **Blue** squares, **Black** circles, and **Red** stars denote diagnostic, overall, and care results.

Patient-level summaries. A concern may appear on only some of a patient's ECGs. In Figure 5a, each concern's maximum score across the patient's recordings is compared with whether that concern is ever recorded. Care macro-AUROC exceeds 0.75 across 8,212 patients, retaining discrimination between patients with and without the corresponding care events. These summaries combine concerns observed at different times, rather than requiring every concern to appear on a single ECG. Their discrimination indicates that *the care component remains informative when assessment is organized around the patient and the available record, not just one recording.*

When a concern changes between ECGs, a responsive score should move in the same direction. We examine 9,375 consecutive within-patient pairs, holding identity fixed while the waveform and labels can change. The measure $\Delta_{t,k} = (p_{t+1,k} - p_{t,k})(y_{t+1,k} - y_{t,k})$ is positive for a score increase accompanying a newly positive label or a decrease accompanying a newly negative one. Direction agreement exceeds 50% for all eight concerns (Figure 5c). It approaches 95% for tachyarrhythmia and exceeds 80% for conduction. High-acuity care reaches roughly 70% agreement, whereas the other care targets remain below 60%. The longitudinal pattern mirrors the difference in overall discrimination: ECG2Act follows rhythm and conduction changes more reliably than care transitions. Even so, agreement above 50% in every target indicates that *care scores capture changes within a patient's ECG record, not just fixed differences in care patterns between patients.*

Dependence on waveform structure. The perturbations in Figure 6 ask whether care prediction depends on the organization of the ECG signal. Phase scrambling reduces diagnostic and care macro-AUROCs by **47.3 and 17.0 percentage points**, respectively, removing most of the original discrimination. Temporal reversal causes smaller but substantial losses of 21.1 and 9.1 points. Since reversal preserves the sample values, the losses show that their temporal arrangement matters to both families. Shifting leads by ± 20 ms reduces the corresponding AUROCs by 7.5 and 1.6 points; care prediction is less affected by these small alignment changes than by phase scrambling or reversal. The temporal dependence is important for interpreting the care results: these targets describe recorded care, not waveform findings, yet their prediction is still sensitive to how ECG samples are arranged. *Care discrimination is therefore not preserved by retaining the sample values alone.*

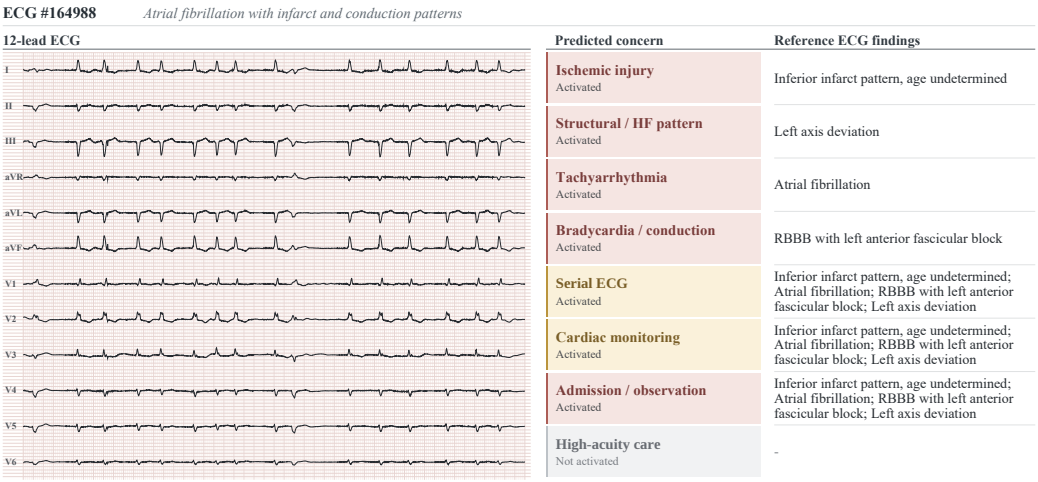


Figure 7: Evidence-based attribution of diagnostic and care-process predictions. An additional inference pass identifies ECG findings associated with each concern. **Red** and **Yellow** denote higher and lower severity among activated concerns; **Gray** denotes non-activation.

4.5 Evidence-Based Attribution

Interpretable attribution through ECG evidence. We use an additional evidence-inference pass to examine whether auxiliary evidence supervision helps explain the predicted concerns (Figure 7). The inferred findings associate ischemic injury with an inferior infarct pattern of undetermined age, tachyarrhythmia with atrial fibrillation, and bradycardia/conduction with right bundle branch block and left anterior fascicular block. The care-process attributions combine these findings with left axis deviation, while their severity grades distinguish the predicted concerns: admission/observation is red, serial ECG and monitoring are yellow, and high-acuity care remains inactive. These correspondences show that *the evidence-supervised model supports qualitative attribution, linking concern predictions to identifiable rhythm, conduction, and morphological features in the ECG.*

5 Discussion

Care information beyond diagnostic labels. ECG2Act retains care discrimination among recordings with identical diagnostic profiles. Reducing these ECGs to four diagnostic labels would discard differences that remain informative about subsequent care. Joint prediction preserves those differences alongside the cardiac problems under assessment, allowing care-process concerns to be evaluated directly from the waveform rather than inferred from a diagnostic summary.

Physiological supervision for prediction and interpretation. Measurement and finding objectives improve both diagnostic and care prediction, showing that the benefits of learning ECG characteristics extend beyond abnormality recognition. Additional evidence inference provides qualitative attribution by linking the predicted concerns to recognizable rhythm, conduction, and morphological findings. The output can therefore present each concern and its severity grade together with the evidence associated with it, rather than leave the reviewer with a numerical score.

6 Conclusion

We introduce ***clinical action concern prediction*** for joint diagnostic and care-process assessment from a single ECG. ECG2Act combines concern-specific queries with physiological supervision, retaining care information beyond the selected diagnostic labels. Additional evidence inference offers qualitative attribution of the predictions to recognizable ECG findings, bringing both concern severity and its associated physiological evidence into the same clinical assessment.

REFERENCES

Arda Altintepe and Kutsev Bengisu Ozyoruk. Serial 12-lead electrocardiogram–based deep-learning model for hospital admission prediction in emergency department cardiac presentations: Retrospective cohort study. *JMIR Cardio*, 9:e80569, 2025. doi: 10.2196/80569.

Rich Caruana. Multitask learning. *Mach. Learn.*, 28:41–75, 1997. doi: 10.1023/A:1007379606734.

Philip de Chazal, Maria O'Dwyer, and Richard B. Reilly. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Trans. Biomed. Eng.*, 51(7): 1196–1206, 2004. doi: 10.1109/TBME.2004.827359.

Brian Gow, Tom Pollard, Larry A. Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W. Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset. *PhysioNet*, 2023. doi: 10.13026/4nqg-sb35. Version 1.0.

Martha Gulati, Phillip D. Levy, Debabrata Mukherjee, et al. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR guideline for the evaluation and diagnosis of chest pain: A report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. *J. Am. Coll. Cardiol.*, 78(22):e187–e285, 2021. doi: 10.1016/j.jacc.2021.07.053.

Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.*, 25:65–69, 2019. doi: 10.1038/s41591-018-0268-3.

Paul A. Heidenreich, Biykem Bozkurt, David Aguilar, et al. 2022 AHA/ACC/HFSA guideline for the management of heart failure: A report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. *J. Am. Coll. Cardiol.*, 79(17): e263–e421, 2022. doi: 10.1016/j.jacc.2021.12.012.

Jiarui Jin, Haoyu Wang, Xingliang Wu, Xiaocheng Fang, Xiang Lan, Zihan Wang, Deyun Zhang, Bo Liu, Yingying Zhang, Xian Wu, Hongyan Li, and Shenda Hong. ECG-R1: Protocol-guided and modality-agnostic MLLM for reliable ECG interpretation. *arXiv preprint arXiv:2602.04279*, 2026. URL https://arxiv.org/abs/2602.04279. Accepted at ICML 2026.

José A. Joglar, Mina K. Chung, Anastasia L. Armbruster, et al. 2023 ACC/AHA/ACCP/HRS guideline for the diagnosis and management of atrial fibrillation: A report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. *J. Am. Coll. Cardiol.*, 83(1):109–279, 2024. doi: 10.1016/j.jacc.2023.08.017.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED. *PhysioNet*, 2023. doi: 10.13026/5ntk-km72. Version 2.2.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, 2024. doi: 10.13026/kpb9-mt58. Version 3.1.

Dani Kiyasseh, Tingting Zhu, and David A. Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *ICML*, volume 139, pp. 5606–5615, 2021. URL https://proceedings.mlr.press/v139/kiyasseh21a.html.

Paul Kligfield, Leonard S. Gettes, James J. Bailey, Rory Childers, Barbara J. Deal, E. William Hancock, Gerard van Herpen, Jan A. Kors, Peter Macfarlane, David M. Mirvis, Olle Pahlm, Pentti Rautaharju, and Galen S. Wagner. Recommendations for the standardization and interpretation of the electrocardiogram: Part I: The electrocardiogram and its technology. *J. Am. Coll. Cardiol.*, 49 (10):1109–1127, 2007. doi: 10.1016/j.jacc.2007.01.024.

Xiang Lan, Feng Wu, Kai He, Qinghao Zhao, Shenda Hong, and Mengling Feng. GEM: Empowering MLLM for grounded ECG understanding with time series and images. In *NeurIPS*, volume 38, 2025. doi: 10.52202/085713-3158. URL https://arxiv.org/abs/2503.06073.

Chin-Sheng Lin, Wei-Ting Liu, Dung-Jang Tsai, et al. AI-enabled electrocardiography alert intervention and all-cause mortality: A pragmatic randomized clinical trial. *Nat. Med.*, 30: 1461–1470, 2024. doi: 10.1038/s41591-024-02961-4.

Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2Label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021. URL https://arxiv.org/abs/2107.10834.

Juan Miguel López Alcaraz, Xicoténcatl López Moran, Erick Dávila Zaragoza, Claas Händel, Richard Koebe, Wilhelm Haverkamp, and Nils Strodthoff. A multimodal deep learning framework for predicting ICU deterioration: Integrating ECG waveforms with clinical data and clinician benchmarking. *arXiv preprint arXiv:2601.06645*, 2026. URL https://arxiv.org/abs/2601.06645.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jay W. Mason, E. William Hancock, and Leonard S. Gettes. Recommendations for the standardization and interpretation of the electrocardiogram: Part II: Electrocardiography diagnostic statement list. *J. Am. Coll. Cardiol.*, 49(10):1128–1135, 2007. doi: 10.1016/j.jacc.2007.01.025.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: An open electrocardiogram foundation model. *JAMIA Open*, 8(5):ooaf122, 2025. doi: 10.1093/jamiaopen/ooaf122.

Sajad Mousavi, Fatemeh Afghah, Abolfazl Razi, and U. Rajendra Acharya. ECGNET: Learning where to attend for detection of atrial fibrillation with deep visual attention. *arXiv preprint arXiv:1812.07422*, 2018. URL https://arxiv.org/abs/1812.07422.

Yoshihiko Raita, Tadahiro Goto, Mohammad Kamal Faridi, David F. M. Brown, Carlos A. Camargo Jr., and Kohei Hasegawa. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit. Care.*, 23:64, 2019. doi: 10.1186/s13054-019-2351-7.

Sunil V. Rao, Michelle L. O'Donoghue, Marc Ruel, et al. 2025 ACC/AHA/ACEP/NAEMSP/SCAI guideline for the management of patients with acute coronary syndromes: A report of the American College of Cardiology/American Heart Association joint committee on clinical practice guidelines. *J. Am. Coll. Cardiol.*, 85(22):2135–2237, 2025. doi: 10.1016/j.jacc.2024.11.009.

Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira Jr., Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat. Commun.*, 11:1760, 2020. doi: 10.1038/s41467-020-15432-4.

Kristin E. Sandau, Marjorie Funk, Andrew Auerbach, Gregory W. Barsness, Kay Blum, Maria Cvach, Rachel Lampert, Jeanine L. May, George M. McDaniel, Marco V. Perez, Sue Sendelbach, Claire E. Sommargren, and Paul J. Wang. Update to practice standards for electrocardiographic monitoring in hospital settings: A scientific statement from the American Heart Association. *Circulation*, 136 (19):e273–e344, 2017. doi: 10.1161/CIR.0000000000000527.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Deyun Zhang, Xiang Lan, Shijia Geng, Qinghao Zhao, Sumei Fan, Mengling Feng, and Shenda Hong. MEETI: A multimodal ECG dataset from MIMIC-IV-ECG with signals, images, features and interpretations. *Sci. Data.*, 13:527, 2026. doi: 10.1038/s41597-026-06796-1.

A Data Sources and Supervision

MEETI and the underlying clinical records. Our ECG corpus is drawn from MEETI (Zhang et al., 2026), a multimodal extension of MIMIC-IV-ECG that aligns waveforms with ECG images, beat-level features, and generated interpretations. We use its record-aligned waveforms as the input to ECG2Act. Its source recordings originate from MIMIC-IV-ECG v1.0 (Gow et al., 2023): ten-second, 12-lead ECGs sampled at 500 Hz at Beth Israel Deaconess Medical Center. The associated MIMIC-IV-ECG machine measurements and reports supply the quantitative and finding supervision described below. MIMIC-IV-ED v2.2 (Johnson et al., 2023) provides emergency-visit times and dispositions; MIMIC-IV v3.1 (Johnson et al., 2024) provides hospital diagnoses, transfers, and intensive-care stays. These resources describe the same patients and encounters, enabling waveform-derived supervision and observed-care labels to be assigned to individual ECGs.

Cohort construction and partitions. We retain MEETI recordings acquired during an emergency-department visit. Subject identifiers and ECG timestamps identify the containing visit; its admission identifier connects the recording to hospital diagnoses and subsequent transfers. Quality screening yields 179,100 ECGs from 82,486 subjects. Patient-level partitioning assigns 143,885 recordings to training, 17,628 to validation, and 17,587 to evaluation. All recordings from a subject remain in one partition. Measurements are standardized using training data, and decision thresholds are selected on validation. The evaluation partition was used for model selection, as reported in Section 4.1.

Sources of auxiliary supervision. The nine measurement targets are *heart rate, RR, PR, QRS duration, QT, corrected QT, and the P, QRS, and T axes*. They are derived from the original machine-measurement fields: rate from RR, intervals from waveform fiducials, and corrected QT with Bazett's formula. Missing measurements have zero regression weight. The fourteen finding targets are extracted from the corresponding machine-generated report lines, using fixed phrase rules, scoped negation, uncertainty handling, and numeric consistency checks. They cover *ST elevation/depression, Q-wave or infarct patterns, atrial fibrillation/flutter, other tachyarrhythmias, rapid rate, bradycardia, atrioventricular block, bundle-branch block, wide QRS, QT prolongation, hypertrophy/structural patterns, low voltage, and axis abnormality*. Confidence weights distinguish reliable from uncertain findings; unavailable evidence is masked. Record identifiers align both auxiliary targets with the MEETI waveforms. Thus, the measurement head learns quantitative ECG physiology, and the finding head learns the report-derived patterns used to construct the diagnostic concerns. Both heads supervise the shared waveform encoder during training.

Four diagnostic concerns. The diagnostic action labels regroup extracted findings into four clinically related problems. **Ischemic/injury** concern combines ST changes, T-wave abnormality, and Q-wave/infarct patterns. **Structural/HF** concern combines hypertrophy/structural patterns, low voltage, and axis abnormality, or a linked-admission heart-failure diagnosis with ICD prefix 428, I50, I110, I130, I132, or I255. **Tachyarrhythmia** combines atrial fibrillation/flutter, other tachyarrhythmias, and rapid rate. **Bradycardia/conduction** combines bradycardia, atrioventricular block, bundle-branch block, and wide QRS. A label is positive when at least one usable constituent is positive. Otherwise it is zero, indicating no positive finding under the extraction rules. T-wave abnormality and admission diagnoses contribute to action labels but are absent from the finding head; QT prolongation is an auxiliary target outside the brady/conduction group.

Four observed-care concerns. **Serial ECG** denotes another ECG from the same patient within the following 24 hours. **Cardiac monitoring** denotes transfer to a monitored unit or an ICU stay. **Admission/observation** uses an admitted ED disposition or observation-unit transfer. **High-acuity care** denotes an ICU stay. Transfer and ICU labels use events beginning from approximately one minute before to 24 hours after the ECG. These labels encode the care processes recorded for each ECG-linked encounter. *Clinical records and auxiliary annotations provide offline supervision for learning all eight concerns; prediction uses the twelve-lead waveform alone.*

B Architecture and Objective Details

Table A1: Joint-lead stem dimensions. Each stage applies convolution, batch normalization, and GELU. Kernel, stride, padding, and output length are measured in temporal samples.

Stage	Input channels	Output channels	Kernel size	Stride / padding	Output length
1	12	128	15	2 / 7	2500
2	128	256	9	2 / 4	1250
3	256	384	7	2 / 3	625
4	384	512	5	4 / 2	157

Joint-lead tokenization. Table A1 expands the tokenization step in *Eq.*(1). A minibatch contains B recordings of shape 12×5000; normalization uses each lead's temporal mean and standard deviation within its own recording. All convolutions use one group, so their filters combine input channels rather than process twelve independent lead streams. The final feature map has 512 channels and 157 temporal positions. Transposing those last two axes produces the $B \times 157 \times 512$ token sequence. Each position thus summarizes a local time region across leads, providing a common source of waveform evidence for both concern families.

Context and readout. The first 157 vectors of a learned 200×512 positional bank are added to the stem output. Six pre-LayerNorm Transformer blocks use eight attention heads of width 64, a 2048-unit feed-forward layer, and dropout 0.1. A final LayerNorm yields $\mathbf{H}$. Eight learned 512-dimensional queries attend to this shared sequence; the residual and normalization in *Eq.*(2) yield $\mathbf{A}$. Each row is mapped to one logit by its own affine head and then passed through a sigmoid. Each query can select overlapping waveform regions, allowing evidence to be shared across the task while leaving the pooling weights specific to the concern being predicted.

Auxiliary supervision. Both auxiliary heads have input and hidden width 512, with **GELU** between their two linear layers. The measurement head outputs nine standardized values from the mean of the encoder tokens. The finding head outputs fourteen logits from the mean of all eight normalized query readouts. Measurement loss consequently supervises the shared encoder, whereas finding loss also reaches the action-query representations. Averaging these readouts does not assign particular findings to particular concerns: the finding target supervises their pooled representation. The concern heads use their waveform readouts directly: measurements and findings supervise the representation during training but are not supplied as inputs when predicting a new ECG.

Missing targets and reductions. Measurement targets use training-split standardization and binary availability masks. Findings use fixed nonnegative observation weights; an unavailable entry contributes zero weight rather than a negative label. For a logit s, binary label y, and positive-class weight α, the elementwise binary loss is defined as

$$\ell_{\mathrm{BCE}}(s, y; \alpha) = -\alpha y \log \sigma(s) - (1 - y) \log(1 - \sigma(s)) \,. \tag{7}$$

Action loss averages over the full $B \times 8$ grid with the capped training-count weights in *Eq.*(4). Finding loss instead sets $\alpha = 1$ and weights entries by availability. Measurement regression applies the following Smooth L1 penalty to each observed target

$$\rho_1(u) = \begin{cases} \frac{1}{2}u^2, & |u| < 1, \\ |u| - \frac{1}{2}, & |u| \geq 1. \end{cases} \tag{8}$$

The denominators in *Eq.*(5) are computed across the entire minibatch: the count of observed measurements and the sum of finding weights, respectively. They are clamped below by one, so an entirely unobserved auxiliary grid contributes zero loss. Targets are not averaged separately before this reduction. The measurement and finding coefficients remain 0.20 and 0.05. After each **AdamW** update, parameter averaging follows $\bar{\theta} \leftarrow 0.995\bar{\theta} + 0.005\theta$, separately from the loss calculation.

C Evaluation Units and Operating Points

Table A2: Concern-specific operating points selected on validation and applied unchanged to evaluation. ***PPV denotes positive predictive value***.

Concern	Threshold (score)	Sensitivity (%)	Specificity (%)	PPV (%)
Ischemic injury	0.48	75.8	89.4	66.2
Structural / HF	0.47	79.5	76.2	66.6
Tachyarrhythmia	0.65	91.3	98.0	92.6
Bradycardia / conduction	0.52	87.6	95.3	84.7
Serial ECG	0.40	75.3	47.3	39.2
Cardiac monitoring	0.38	83.0	32.6	39.7
Admission / observation	0.24	94.6	20.7	67.9
High-acuity care	0.29	58.0	84.1	32.0

Table A3: Evaluation weighting and binary prediction-set summaries. The patient-equal result reweights individual ECGs; it does not use the maximum-score patient aggregation in Figure 5a.

Quantity	Value
Recording-weighted macro-AUROC (17,587 ECGs)	0.825
Patient-equal macro-AUROC (8,212 patients)	0.836
Nominal marginal prediction-set coverage	0.900
Observed micro-averaged prediction-set coverage	0.899
Mean binary prediction-set size	1.261

Cohort and analysis units. The cohort contains 179,100 ECGs from 82,486 subjects, with 143,885 training, 17,628 validation, and 17,587 evaluation recordings. Patients do not cross partitions. Evaluation includes 8,212 patients and 9,375 consecutive within-patient pairs; these are different views of the same recordings rather than additional cohorts. Each model prediction uses one waveform. Because this partition informed model selection, patient separation does not make the reported estimates independent of that selection process.

Operating-point interpretation. The family and overall macro scores in Table 2 weight the relevant four or eight concerns equally. Table A2 characterizes the resulting thresholded outputs, rather than selecting new thresholds on evaluation. Sensitivity is measured among positive cases, specificity among negative cases, and PPV among flagged cases. Admission/observation trades limited specificity for high sensitivity, while the high-acuity operating point is more specific but misses more positive cases. These are target-specific operating characteristics; neither a higher F1 nor a higher AUROC by itself fixes the balance between missed and additional flags.

Aggregation and reweighting. For patient u with recording index set I_u, aggregation uses $\max_{i \in I_u} p_{ik}$ as the score and $\max_{i \in I_u} y_{ik}$ as the label for concern k (Figure 5a). Different concerns may attain their maxima on different ECGs; the aggregated vector need not describe any single recording. Patient-equal weighting instead retains every recording-level score and label, giving recording i weight proportional to $1/|I_u|$. The change in Table A3 therefore reflects whose records contribute to the metric, not a revised model. Aggregation changes the score and target summaries; reweighting changes the population weighting of the original task.

Binary prediction sets. The supplementary split-conformal summary uses nominal miscoverage $\epsilon = 0.10$. Observed micro-averaged coverage is 0.899, with mean set size 1.261 binary classes per ECG–concern instance (Table A3). Each set contains candidate values for one binary label, not a selection among eight concerns. Coverage asks whether that observed label is included; set size measures how much ambiguity remains. These summaries do not establish calibrated probabilities, subgroup-conditional coverage, or simultaneous coverage of all outputs. Within-patient dependence and evaluation-informed selection limit the interpretation of the nominal coverage level.

D AUXILIARY-SUPERVISION CONFIGURATIONS

Controlled supervision settings. Table 3 varies the two physiological objectives while retaining action-concern supervision. Let $u_M, u_E \in \{0, 1\}$ indicate whether measurement and finding supervision are active. The objective becomes

$$\mathcal{L}(u_M, u_E) = \mathcal{L}_{\text{act}} + 0.20 u_M \mathcal{L}_{\text{meas}} + 0.05 u_E \mathcal{L}_{\text{evid}}.$$

The four settings are both objectives off, findings only, measurements only, and both on. The reported ablation uses an NVIDIA RTX 4090 GPU. An enabled branch keeps the routing, coefficient, and reduction defined in Section 3; a disabled branch contributes no auxiliary loss. The switches thus change supervision during training, not the availability of clinical information at evaluation. All configurations predict the final concerns from a waveform.

Contribution beyond action labels. With both switches off, the model still learns the complete joint target space. This is the appropriate reference for asking whether physiological targets contribute beyond action-label supervision. Both single-objective settings exceed that reference in diagnostic and care family AUROC. The care-side gains are important to the task formulation: measurement and finding targets describe ECG properties rather than additional care events. Their contribution is consistent with physiological supervision improving a representation used by both families, even though the families have different prediction targets. It does not depend on providing measurements or reference findings as extra inputs when evaluating the model.

Standalone and conditional effects. Measurements alone outperform findings alone in both family AUROCs, but this comparison does not test whether findings remain useful once measurements are available. Enabling findings on top of measurements raises diagnostic AUROC from 0.9325 to 0.9377 and care AUROC from 0.7077 to 0.7133. The combined setting also exceeds findings alone. These conditional comparisons, rather than the standalone ordering, motivate retaining both objectives. The incremental gains support using both signals in this configuration, without implying equal contributions, additive effects, or a particular physiological mechanism behind the improvements.

Agreement across metrics. The combined configuration also has the highest macro-AUPRC and macro-F1 in Table 3. These summaries emphasize precision–recall and thresholded prediction rather than ranking alone. Their agreement with the family AUROCs makes the preference for combined supervision less dependent on a single reported metric. This remains a comparison of the four configurations, not a claim of statistical significance or superiority at every operating threshold. The concern-wise results in Table 2 describe a separate evaluation summary and should not be used to reconstruct configuration-level results from rounded endpoint values.

What is held fixed. The measurement head reads pooled encoder tokens, while the finding head reads pooled action-query representations (Figure 2). Turning the corresponding loss on allows it to supervise those representations and their upstream parameters. This routing explains the two levels of physiological guidance, but it is not itself varied by the switches. The experiment therefore evaluates the presence of each objective within the specified architecture. It cannot attribute the gain specifically to the pooling rule, query design, or attachment point without comparisons that vary those choices. The conclusion concerns supervision in the joint predictor, rather than a unique mechanism for converting ECG physiology into care scores.

Relation to the joint task. The supervision study and the conditional care-information analyses address complementary parts of the main narrative. Table 3 asks whether explicit ECG-property targets help learn the broader prediction task. The probes and fixed-profile comparisons in Section 4.3 ask whether its care outputs retain information that the four diagnostic labels leave unresolved. Better training performance alone would not establish this additional information; within-profile ranking alone would not identify which training objective helped obtain it. Together they connect the proposed learning design to two distinct empirical questions: how to supervise the shared representation and what information its separate care outputs preserve.

E Additional Care-Information Analyses

Table A4: Care-event concentration at three selection budgets. Capture is the percentage of all positive cases selected for that concern; lift compares capture with the expected value under random selection.

Concern	Top 1%		Top 10%		Top 20%	
	Capture	Lift	Capture	Lift	Capture	Lift
Serial ECG	2.70	2.70	20.58	2.06	34.93	1.75
Cardiac monitoring	2.38	2.38	18.45	1.84	31.76	1.59
Admission / observation	1.38	1.38	14.00	1.40	27.05	1.35
High-acuity care	6.01	6.01	37.16	3.72	57.13	2.86

From probe augmentation to fixed profiles. Figure 4a compares validation-fitted care probes using the four diagnostic labels alone or augmented with ECG-derived scores. This tests incremental prediction relative to a particular diagnostic representation and fitted probe. Table 4 then fixes the entire four-label vector within each comparison group. A label-only predictor must tie the cases in such a group, irrespective of its functional form. The second analysis therefore tests an explanation that the first comparison cannot exclude by itself: whether an apparent gain reflects only the limitations of the fitted label-only probe. It tests whether recordings with identical diagnostic summaries can still be distinguished by their care scores when their observed care labels differ.

What the within-profile average measures. The thirteen reported profiles cover 17,203 of the 17,587 evaluation ECGs. For a given care target, each within-profile AUROC compares positive and negative cases drawn from the same diagnostic profile. A tie receives half credit, so a score that is constant within a profile has no discriminatory ordering there. The final row of Table 4 averages these AUROCs with the listed ECG counts as weights. It does not pool all recordings: a pooled AUROC would also compare cases from different diagnostic profiles. The result isolates ranking beyond those label differences, while allowing disease severity, waveform morphology, and other clinical context to vary within each group.

Neighborhood label disagreement. A supplementary neighborhood analysis considers 5,000 query ECGs. For each query, it examines up to eleven neighbors for a recording with the same diagnostic profile but a different care-label vector. Such a neighbor is found for 4,448 queries (89.0%); the search does not find one for the remaining 552 (11.0%). This shows that diagnostic agreement and care disagreement coexist in the inspected neighborhoods. It does not estimate their frequency throughout the population: a successful search establishes one qualifying match, while failure within the search limit does not exclude a match elsewhere. The statistic supplements, but does not replace, the score-based conditional discrimination test.

Capture and lift at a review budget. For a care concern, let P denote its positive cases and S_q the recordings selected by the highest scores at budget fraction q. We summarize this ranking by

$$\mathrm{Capture}(q) = \frac{|S_q \cap P|}{|P|}, \qquad \mathrm{Lift}(q) = \frac{\mathrm{Capture}(q)}{q}.$$

Capture uses all positive cases as its denominator; precision would use the selected set $|S_q|$. Expected capture under random selection equals the selection fraction, which supplies the denominator of lift. Table A4 expresses capture as a percentage and lift as a dimensionless ratio. The four care scores are ranked independently, so their selected sets need not contain the same ECGs.

Concentration versus coverage. For high-acuity care, increasing the budget from 1% to 10% and 20% raises capture from 6.01% to 37.16% and 57.13%, while lift falls from 6.01 to 3.72 and 2.86. A narrower selection is more enriched; a broader selection reaches more positive cases. The budgets in Table A4 extend the top-5% view in Figure 4b and make this concentration–coverage trade-off explicit. They do not impose one threshold on all care concerns. This is a retrospective ordering of observed events, not a measurement of whether acting on the selected cases would improve care.

F LONGITUDINAL AND PERTURBATION DETAILS

Table A5: Aligned score changes on label-transition pairs, in sigmoid-score percentage points. Counts refer to changes in the indicated concern, not distinct patients.

Concern	Label transitions (pairs)	Mean aligned change (percentage points)
Ischemic injury	1779	26.26
Structural / HF	2075	15.41
Tachyarrhythmia	1758	75.18
Bradycardia / conduction	1508	51.49
Serial ECG	4201	2.44
Cardiac monitoring	2381	3.23
Admission / observation	2257	3.33
High-acuity care	927	13.65

Table A6: Intact-segment order control. One-second segments retain their internal samples but change order. Values are perturbed minus original AUROC, in percentage points.

Perturbation	Overall	Diagnostic	Care
One-second segment permutation	−0.361	−0.544	−0.178

Two uses of consecutive recordings. The comparisons in Figure 5b and Figure 5c use consecutive ECGs differently. In panel b, a current score is evaluated against the current label and then the next-recording label; the prediction itself is held fixed. The transition analysis instead recomputes a score for each recording and measures the change between them. The former characterizes the association of one recording with two label times; the latter characterizes responsiveness to an observed label transition. Neither comparison defines a uniform time interval, so the results do not constitute a fixed-horizon forecasting benchmark.

Response magnitude and direction. For concern k changing label between recordings t and $t+1$, the score difference is aligned with the observed label change:

$$\Delta_{t,k} = (p_{t+1,k} - p_{t,k})(y_{t+1,k} - y_{t,k}), \qquad y_{t+1,k} \neq y_{t,k}. \tag{9}$$

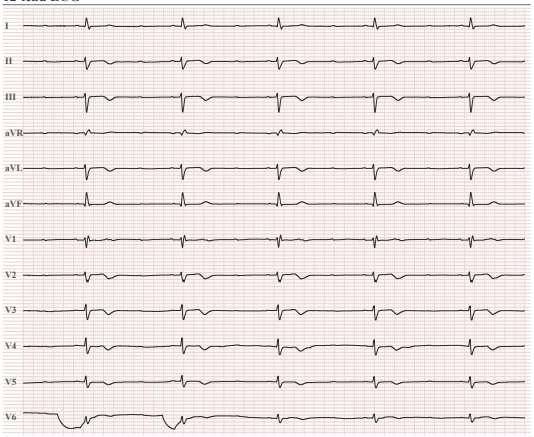
A change from label 0 to 1 is aligned with a score increase; a change from 1 to 0 is aligned with a decrease. Table A5 reports 100 times the mean aligned change, in sigmoid-score percentage points rather than AUROC points. Direction agreement in Figure 5c is the proportion with $\Delta_{t,k} > 0$. The mean records the size of the response, while the proportion records how consistently its direction agrees with the label. Transition counts differ because each concern changes on its own subset of the 9,375 pairs; the rows do not partition the patient population.

Intact segments as a control. The perturbations in Figure 6 alter the waveform supplied to the same model. Phase scrambling, temporal reversal, and inter-lead skew affect different aspects of waveform organization. The additional control in Table A6 changes the order of one-second segments while retaining each segment's internal samples. Its small AUROC changes contrast with the larger phase and reversal effects. The evidence therefore distinguishes intact-segment reordering from the other tested transformations: temporal structure cannot be reduced to a single undifferentiated property. The comparisons characterize these specified inputs, not invariance at other perturbation strengths.

Scope of the behavioral evidence. Diagnostic and care outputs show different transition magnitudes and perturbation responses despite sharing an encoder. This motivates inspecting each target family rather than relying on the overall mean. A smaller absolute care-AUROC drop does not establish greater general robustness, because the families start at different performance levels and the perturbations lack a common severity scale. Together, the longitudinal and input controls describe how the joint output varies with recordings. They complement the conditional care-information analysis without identifying the physiological cause of an individual prediction.

G Additional Reference-Annotated Examples

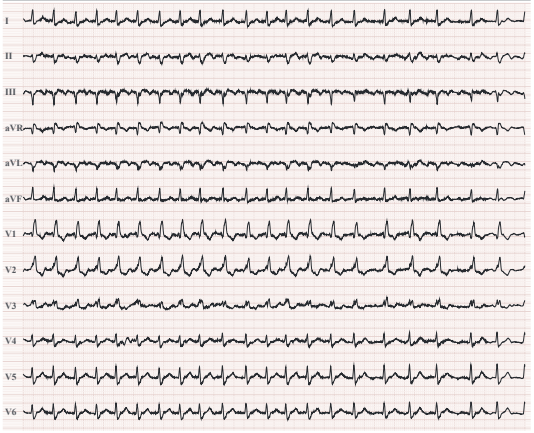
ECG #131087 *AV block with structural and conduction findings*



Predicted concern	Reference ECG findings
Ischemic injury Not activated	-
Structural / HF pattern Activated	LVH with IVCD; Left axis deviation
Tachyarrhythmia Not activated	-
Bradycardia / conduction Activated	3:1 AV block; Wide QRS (152 ms)
Repeat ECG Activated	3:1 AV block; Wide QRS (152 ms); LVH with IVCD; Left axis deviation
Cardiac monitoring Activated	3:1 AV block; Wide QRS (152 ms); LVH with IVCD; Left axis deviation
Admission / observation Activated	3:1 AV block; Wide QRS (152 ms); LVH with IVCD; Left axis deviation
High-acuity review Activated	3:1 AV block; Wide QRS (152 ms); LVH with IVCD; Left axis deviation

Figure A1: ECG #131087: AV block and structural findings; all four care concerns are active.

ECG #47399 *Rapid rhythm with infarct and conduction patterns*



Predicted concern	Reference ECG findings
Ischemic injury Activated	Inferior infarct pattern, age undetermined
Structural / HF pattern Activated	Left axis deviation
Tachyarrhythmia Activated	Sinus tachycardia (129 bpm)
Bradycardia / conduction Activated	RBBB with left anterior fascicular block
Repeat ECG Activated	Inferior infarct pattern, age undetermined; Sinus tachycardia (129 bpm); RBBB with left anterior fascicular block; Left axis deviation
Cardiac monitoring Activated	Inferior infarct pattern, age undetermined; Sinus tachycardia (129 bpm); RBBB with left anterior fascicular block; Left axis deviation
Admission / observation Activated	Inferior infarct pattern, age undetermined; Sinus tachycardia (129 bpm); RBBB with left anterior fascicular block; Left axis deviation
High-acuity review Not activated	-

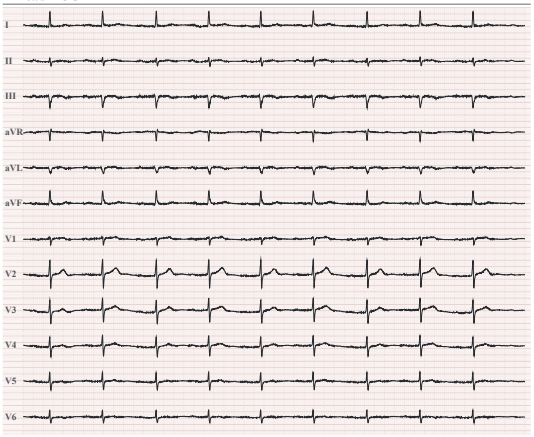
Figure A2: ECG #47399: rapid rhythm with infarct and conduction patterns; high acuity is inactive.

Interpreting the conduction concern. Figures A1 and A2 extend the evidence attribution in Figure 7 to different conduction abnormalities. Additional evidence inference identifies 3:1 AV block with QRS widening (152 ms) in the first case, and right bundle branch block with left anterior fascicular block in the second. Both activate the same concern, but each attribution names a specific conduction abnormality that the reviewer can compare with the corresponding ECG waveform.

Graded care priorities. The two recordings also differ in their predicted care severity. Figure A1 has two diagnostic flags but grades all four care concerns red (higher severity). Figure A2 has four diagnostic flags, yet high-acuity care is gray (inactive), monitoring and admission/observation are yellow (lower severity), and only serial ECG is red. The grades describe the concern assigned to each care process, not a severity score obtained by counting diagnostic flags.

Care within a diagnostic profile. A closer comparison comes from Figures A4 and A5, which share structural/HF and conduction flags. Their inferred evidence differs: broad-QRS abnormalities in the former, low voltage and QT prolongation in the latter. Both activate monitoring and admission/observation, but only the broad-QRS case flags serial ECG.

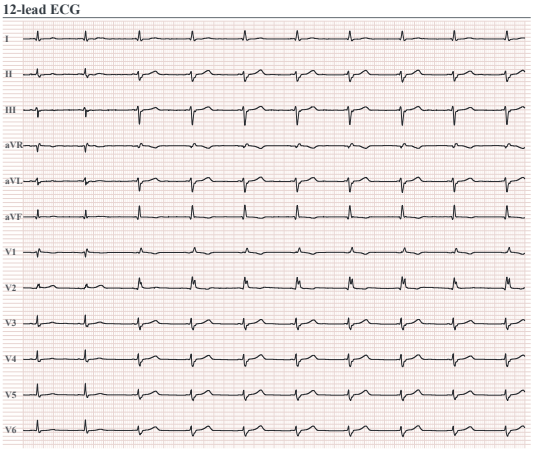
ECG #98750 *Bradycardia with AV block and infarct pattern*



Predicted concern	Reference ECG findings
Ischemic injury Activated	Inferior infarct pattern, age undetermined
Structural / HF pattern Activated	Left axis deviation
Tachyarrhythmia Not activated	-
Bradycardia / conduction Activated	Sinus bradycardia (56 bpm); 1st-degree AV block (PR 324 ms)
Repeat ECG Activated	Inferior infarct pattern, age undetermined; Sinus bradycardia (56 bpm); 1st-degree AV block (PR 324 ms); Left axis deviation
Cardiac monitoring Activated	Inferior infarct pattern, age undetermined; Sinus bradycardia (56 bpm); 1st-degree AV block (PR 324 ms); Left axis deviation
Admission / observation Activated	Inferior infarct pattern, age undetermined; Sinus bradycardia (56 bpm); 1st-degree AV block (PR 324 ms); Left axis deviation
High-acuity review Not activated	-

Figure A3: ECG #98750: bradycardia with AV block and an infarct pattern.

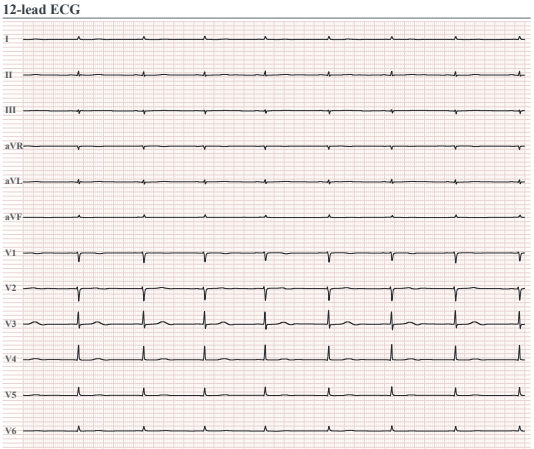
ECG #95160 *Bradycardia with broad-QRS structural pattern*



Predicted concern	Reference ECG findings
Ischemic injury Not activated	-
Structural / HF pattern Activated	LVH with IVCD
Tachyarrhythmia Not activated	-
Bradycardia / conduction Activated	Sinus bradycardia (57 bpm); Right bundle branch block; Wide QRS (138 ms)
Repeat ECG Activated	Sinus bradycardia (57 bpm); Right bundle branch block; Wide QRS (138 ms); LVH with IVCD
Cardiac monitoring Activated	Sinus bradycardia (57 bpm); Right bundle branch block; Wide QRS (138 ms); LVH with IVCD
Admission / observation Activated	Sinus bradycardia (57 bpm); Right bundle branch block; Wide QRS (138 ms); LVH with IVCD
High-acuity review Not activated	-

Figure A4: ECG #95160: bradycardia with a broad-QRS structural pattern.

ECG #42772 *Bradycardia with low voltage and prolonged QT*



Predicted concern	Reference ECG findings
Ischemic injury Not activated	-
Structural / HF pattern Activated	Low limb-lead voltage
Tachyarrhythmia Not activated	-
Bradycardia / conduction Activated	Sinus bradycardia (48 bpm)
Repeat ECG Not activated	-
Cardiac monitoring Activated	Sinus bradycardia (48 bpm); Prolonged QT interval; Low limb-lead voltage
Admission / observation Activated	Sinus bradycardia (48 bpm); Prolonged QT interval; Low limb-lead voltage
High-acuity review Not activated	-

Figure A5: ECG #42772: low-voltage bradycardia with prolonged QT.

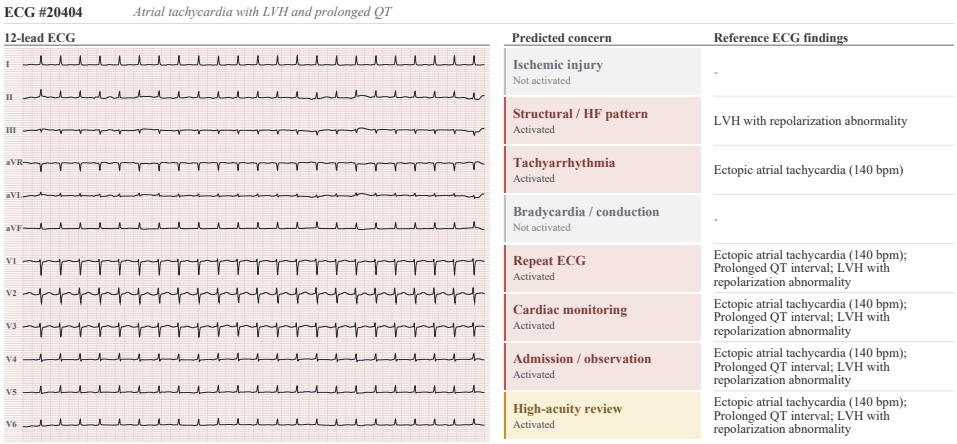


Figure A6: ECG #20404: atrial tachycardia with LVH and prolonged QT.

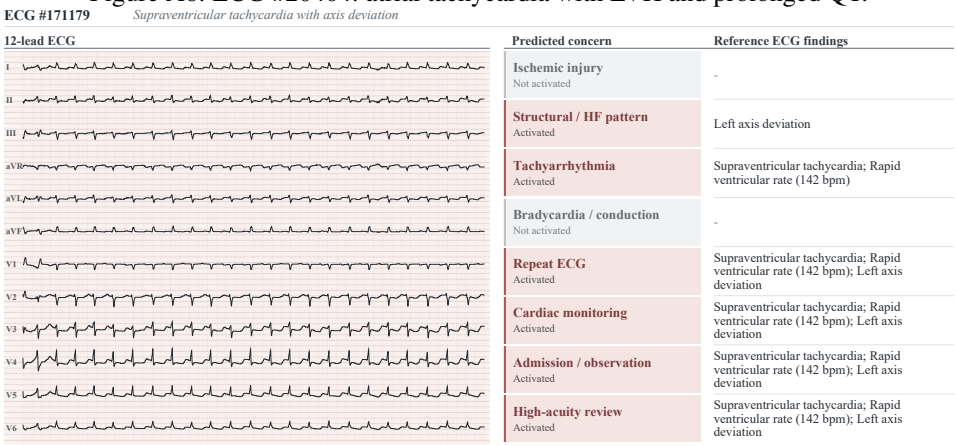


Figure A7: ECG #171179: supraventricular tachycardia with axis deviation.

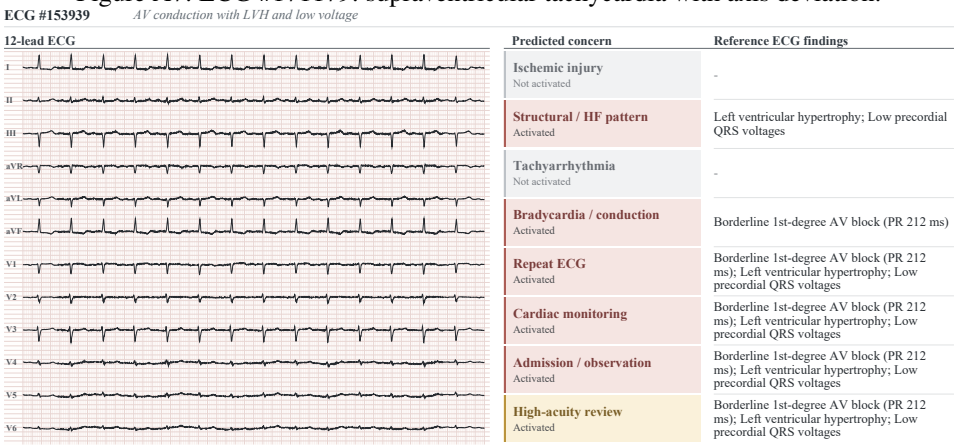


Figure A8: ECG #153939: AV conduction with LVH and low voltage.

AI USE STATEMENT

Large language models, including Claude Code, were used only for AI-assisted drafting, editing, and literature checks. They did not influence experimental design, execution, analysis, or interpretation. All scientific decisions, analyses, and results were produced and fully verified solely by the authors.