Shortcut-Aware Modeling of Serial ECGs for Cardiac Sarcoidosis

Humza A. Iqbal HIQBAL@OTTAWAHEART.CA
Telfer School of Management, University of Ottawa
University of Ottawa Heart Institute
Ottawa, Ontario, Canada

David Birnie DBIRNIE@OTTAWAHEART.CA
Division of Cardiology, Department of Medicine, University of Ottawa
University of Ottawa Heart Institute
Ottawa, Ontario, Canada

Christopher L. F. Sun SUN@TELFER.UOTTAWA.CA
Telfer School of Management, University of Ottawa
University of Ottawa Heart Institute
Ottawa, Ontario, Canada

Abstract

Cardiac sarcoidosis is difficult to diagnose, and serial ECG data may encode both physiologic signal and observation-process shortcuts. Here, observation-process shortcuts are predictive signals from ECG count and inter-ECG intervals rather than ECG waveform physiology. We study patient-level prediction of future CS diagnosis from serial ECG data collected before clinical presentation at a tertiary cardiac center. We compare single-ECG models, shortcut-only controls, shortcut-vulnerable serial ECG models, and shortcut-aware serial ECG models under unmatched evaluation and matched shortcut-control evaluation. We introduce RECAP, a shortcut-aware joint architecture-and-training design that represents prior ECGs relative to the patient's latest eligible ECG, conditions these relationships on elapsed time, and uses matched-pair loss to reduce reliance on the measured shortcut variables. In the unmatched evaluation, shortcut-only controls are strongly predictive, showing that apparent serial gains can be driven by shortcut variables. In the matched shortcut-control evaluation, shortcut-only controls collapse to chance, and RECAP achieves the highest matched AUROC among the evaluated configurations. These results show that serial ECG waveforms retain predictive signal after targeted shortcut control and that rigorous evaluation in serial ECG settings requires explicit shortcut-only controls and matched shortcut-control evaluation.

1. Introduction

Cardiac sarcoidosis (CS) is an uncommon but potentially fatal manifestation of systemic sarcoidosis. CS is an inflammatory cardiomyopathy characterized by non-necrotizing granulomatous inflammation of the myocardium. Inflammation and subsequent scarring can lead to high-grade atrioventricular block, ventricular arrhythmias, ventricular dysfunction, heart failure, or sudden cardiac death. Early CS recognition is critical because it changes clinical decisions regarding immunosuppression, rhythm surveillance, device therapy, and advanced

imaging follow-up (Cheng et al., 2024; Lehtonen et al., 2023). However, CS remains difficult to diagnose because myocardial infiltration is often spatially heterogeneous, which limits the sensitivity of endomyocardial biopsy, and may remain clinically silent. Indeed, CS is often identified by integrating comprehensive clinical findings, extracardiac histology when available, and advanced imaging rather than by a single definitive test (Cheng et al., 2024; De Bortoli et al., 2024; Lehtonen et al., 2023). This combination of high consequence, low prevalence, and costly diagnostics makes earlier case finding attractive but difficult.

Within this diagnostic pathway, the 12-lead ECG is a natural screening candidate because it is inexpensive, ubiquitous, and repeatedly collected in routine cardiovascular care, and because CS can be visible through conduction and repolarization abnormalities. Indeed, early disease-specific studies suggest that informative signal can be recovered from single ECGs and routine cardiovascular data (Katsushika et al., 2022; de Melo Jr et al., 2025). However, a single tracing cannot show whether an abnormality is new, progressive, or stable relative to that patient's own baseline. That limitation matters because CS risk assessment depends on whether ECG findings are newly emerged, worsening, or stable over time, and a single ECG cannot provide that patient-specific temporal context (Cheng et al., 2024; Lehtonen et al., 2023). Serial ECG data could therefore make borderline findings more informative and reduce dependence on observing one ECG at the most revealing point in the disease course (Sbrollini et al., 2019; Choi et al., 2024).

Although serial ECGs can improve prediction by providing patient-specific temporal context, they can also create opportunities for shortcut learning. ECG records are irregular and clinically triggered rather than sampled on a fixed schedule. Patients with symptoms, abnormal prior tests, referral to specialty care, or closer surveillance contribute different numbers of ECGs and different inter-ECG-interval patterns. ECG count and inter-ECG intervals are therefore shortcut variables rather than the clinical signal of interest. Work on informative observation, presentness, and undertesting shows that such variables can be predictive on their own, including through non-clinical differences in access to care and follow-up intensity (Weiskopf et al., 2013; Lipton et al., 2016; Fauber and Shelton, 2018; Chang et al., 2022; Ong Ly et al., 2024). Because access to care is itself unequally distributed, this shortcut pathway can also reinforce health-equity distortions rather than only inflate discrimination. It is therefore critical to test whether serial ECG models are vulnerable to ECG count and inter-ECG intervals, and whether shortcut-aware evaluation and model design can mitigate that vulnerability while preserving predictive waveform signal.

Contributions. First, we show that ECG count and inter-ECG intervals provide strong shortcut signals in serial ECG prediction by evaluating shortcut-only controls before and after matching on these variables. Second, we develop a shortcut-control evaluation protocol that tests whether waveform-based models retain predictive signal after cases and controls are matched on the measured shortcut variables. Third, we introduce RECAP, a joint architecture-and-training design that represents prior ECGs relative to the anchor ECG, conditions these relations on elapsed time, and trains with a matched-pair objective. RECAP achieves the highest matched AUROC estimate among the evaluated configurations, and its component analysis shows stepwise gains under matched-pair training.

Generalizable Insights about Machine Learning in the Context of Healthcare

- Serial ECGs can make the latest ECG more informative by capturing patient-specific changes over time, supporting ECG-based prediction models.
- Serial healthcare data can mix physiologic signal with observation-process variables because the observation process, including who is tested, how often, and how closely patients are followed, can itself be predictive, creating opportunities for shortcut learning.
- Shortcut-only control models and shortcut-learning evaluations are needed to test whether model inference relies on clinical signals rather than shortcut variables.

2. Related Work

2.1. ML for Cardiac Sarcoidosis

Disease-specific ML in cardiac sarcoidosis remains limited and spans different modalities and endpoints. Prior studies have used single 12-lead ECGs for CS detection, echocardiographic movies for CS classification, and multimodal imaging-plus-clinical data for sudden cardiac death risk stratification (de Melo Jr et al., 2025; Katsushika et al., 2022; Lai et al., 2026). These studies support the plausibility of disease-specific ML in CS, but they address different tasks from ours. They do not directly address prediction of future CS diagnosis from serial ECG data before clinical presentation.

2.2. Serial ECG Modeling

Outside cardiac sarcoidosis, serial ECG studies have used patient-specific comparison, engineered difference features, and deep learning over repeated 12-lead ECGs (Sunemark et al., 1998; Gregg et al., 2012; Sbrollini et al., 2019, 2023; Bouzid et al., 2023; Choi et al., 2024; Nishihara et al., 2025). These studies support the idea that prior ECGs can improve interpretation or prediction. They were evaluated under the observed testing process, however, rather than separating waveform signal from shortcut variables such as ECG count and inter-ECG intervals. Prior serial ECG studies have not directly tested whether serial ECG models are vulnerable to those shortcut variables and shortcut learning pathways.

2.3. Shortcut Learning in Healthcare ML

Observation-process bias is well documented in healthcare ML. Measurement presence, timing, ordering, and record completeness can reflect illness severity, diagnostic workup, and non-clinical differences in access to care or follow-up intensity, making them predictive apart from the intended physiologic signal (Weiskopf et al., 2013; Sisk et al., 2021; Lipton et al., 2016; Fauber and Shelton, 2018; Che et al., 2018; Agniel et al., 2018; Tan et al., 2023; Chang et al., 2022). Related work addresses these risks through descriptive stress tests, evaluation designs that restrict or match on shortcut variables, and model-side debiasing or invariance-promoting training (Futoma et al., 2021; Brown et al., 2023; Ong Ly et al., 2024; Jabbour et al., 2020; Marcinkevics et al., 2022). This paper builds on those ideas by pairing matched evaluation with model-side controls in a serial ECG setting. The literature does

not yet answer whether serial ECG models retain predictive value for future CS diagnosis after shortcut control based on ECG count and inter-ECG intervals.

3. Methods

3.1. Problem Formulation

We study patient-level prediction of future CS diagnosis from serial ECG data collected before clinical presentation (Figure 1).

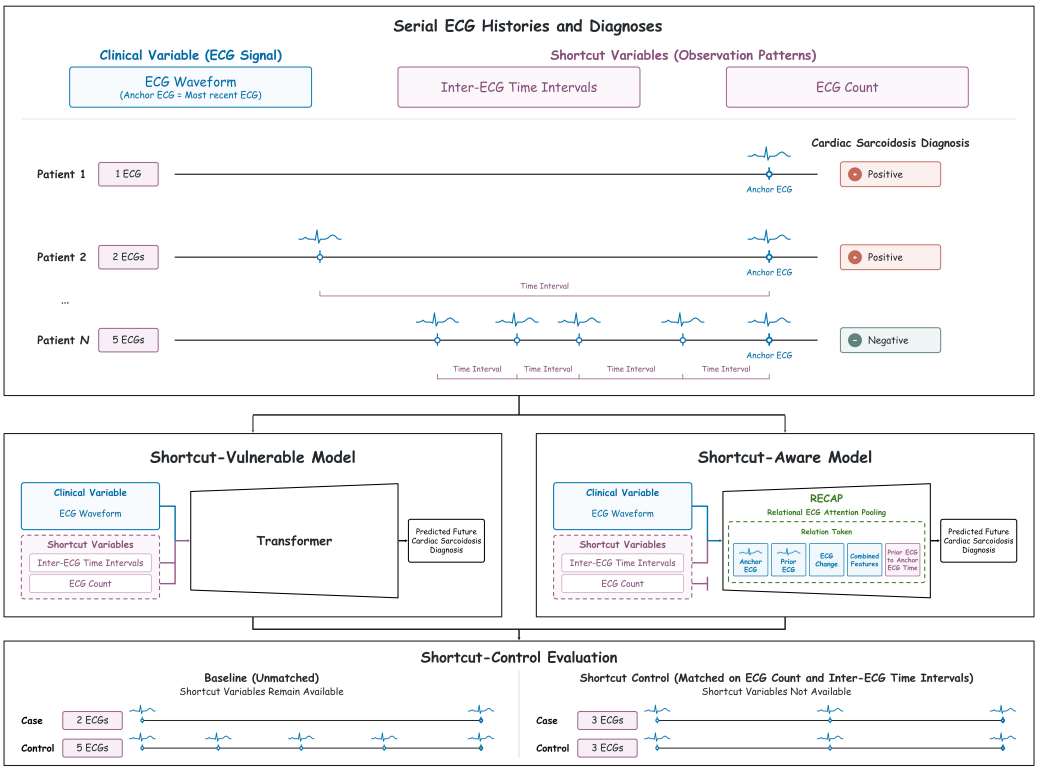

Figure 1: Overview of the study design. The top panel separates ECG waveform information from shortcut variables defined by ECG count and inter-ECG intervals. The middle panel contrasts shortcut-vulnerable serial modeling with RECAP, which combines anchor–prior relational attention and matched-pair training. The bottom panel compares unmatched and shortcut-matched evaluation.

Each patient, indexed by i, contributes a binary outcome, with $y_i = 1$ if CS diagnosis is recorded after the anchor ECG, defined as the patient's latest eligible ECG and the ECG used by the model to make its prediction, and $y_i = 0$ otherwise, together with zero or more earlier eligible ECGs. The superscript anc denotes quantities associated with this anchor ECG.

Let the eligible pre-presentation ECG history be

$$\mathcal{H}_i = \{(x_{ij}, t_{ij})\}_{j=1}^{c_i}, \qquad t_{i1} < \cdots < t_{ic_i}, \tag{1}$$

where x_{ij} is the waveform of ECG j, t_{ij} is its acquisition time, and c_i is the number of eligible ECGs. The index $j = 1, \ldots, c_i$ orders this patient's eligible ECGs chronologically, with $j = 1$ oldest and $j = c_i$ newest, and $\mathcal{H}_i$ denotes that complete ordered history. The anchor waveform and anchor time are therefore $x_i^{\text{anc}} = x_{i,c_i}$ and $t_i^{\text{anc}} = t_{i,c_i}$. For CS-positive patients, only ECGs recorded strictly before the documented presentation date are eligible. For patients without documented CS, only ECGs recorded before administrative end of follow-up are eligible. The task is to predict y_i from $\mathcal{H}_i$ using only information available up to the anchor boundary. The label represents future CS diagnosis rather than independently confirmed disease onset.

3.2. Why Serial ECG Models Are Shortcut-Vulnerable

The main shortcut risk in this study comes from ECG count and inter-ECG intervals rather than from ECG waveform information. Routine serial ECG data are sparse and irregular, and the observation process is shaped by both clinical and non-clinical factors. Symptoms, prior abnormalities, and referral for further cardiac workup can reasonably lead to more frequent ECGs, but differences in access to care and other non-clinical factors can also change how often patients are observed. Patients with greater access to care may therefore contribute more ECGs and shorter inter-ECG intervals, which can create misleading discrimination from the observation process rather than from clinically informative ECG waveforms. In this paper, the primary shortcut variables are the ECG count c_i and the consecutive inter-ECG intervals

$$\delta_{ij} = t_{ij} - t_{i,j-1}, \quad j = 2, \ldots, c_i. \tag{2}$$

Prior work on missingness, presentness, undertesting, and shortcut learning makes it plausible that such observation patterns can be predictive on their own (Lipton et al., 2016; Fauber and Shelton, 2018; Chang et al., 2022; Ong Ly et al., 2024). We therefore treat c_i and δ_{ij} as shortcut variables rather than the clinical signal of interest. They may still correlate with future CS diagnosis through clinically appropriate workup as well as non-clinical differences in care access, which is exactly why they can produce misleading discrimination.

This shortcut pathway is visible in the study data themselves. In the unmatched evaluation cohort, CS-positive patients and controls differ substantially in ECG count and inter-ECG-interval distributions (Figure 2).

3.3. Shortcut-Aware Serial Modeling

3.3.1. Relational ECG Attention Pooling

Here we present relational ECG attention pooling (RECAP), a joint shortcut-aware serial ECG architecture and training design that forms the paper's main methodological contribution. Its architecture aims to use prior ECG waveforms without supplying ECG count

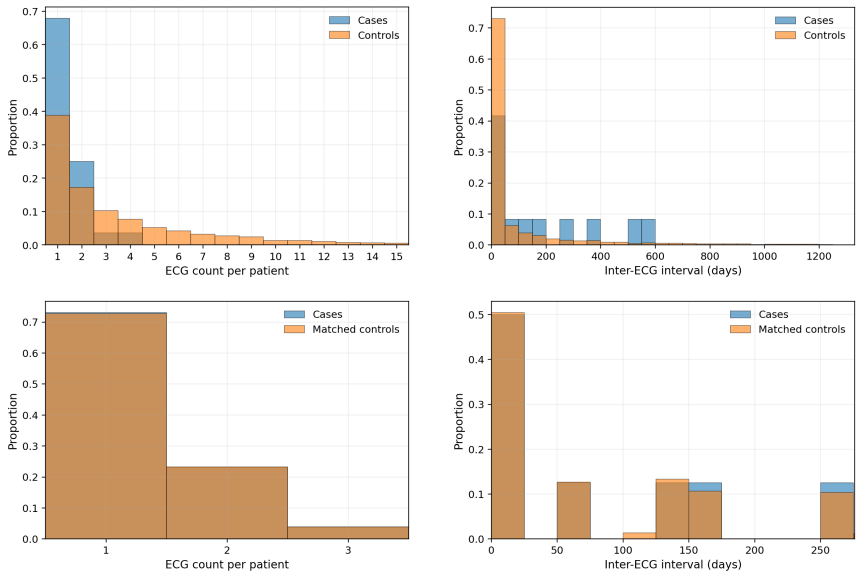

Figure 2: Distributions of ECG count and inter-ECG intervals before and after shortcut matching. Left panels show ECG count per patient, and right panels show inter-ECG intervals. Top panels compare CS-positive patients with unmatched controls, whereas bottom panels compare CS-positive patients with matched controls. Bars show within-group proportions.

or consecutive inter-ECG intervals as separate scalar features. This distinction matters because elapsed time is not intrinsically spurious. When interpreted together with waveform change, it can indicate how quickly ECG morphology evolves. Accordingly, the architecture conditions each anchor–prior waveform relation on its elapsed time rather than supplying elapsed time directly to the prediction heads. This allows the model to interpret waveform change relative to the time over which it occurred, while avoiding a direct time-only pathway to the classifier.

For patient i, j indexes each prior ECG. A frozen pretrained ECGFounder encoder (Li et al., 2025) produces a 1,024-dimensional embedding for each ECG. A shared trainable projection block maps every anchor and prior embedding into the same 256-dimensional task-specific space, yielding z_{ij} for prior ECG j and z_i^{anc} for the anchor ECG. This provides task-specific adaptation without updating ECGFounder and places anchor and prior ECGs in the same feature space before their differences and products are computed. For each anchor–prior pair, we define standardized elapsed-time scalar s_{ij} and relation token r_{ij} as follows.

$$\begin{aligned} d_{ij} &= t_i^{\text{anc}} - t_{ij}, \\ s_{ij} &= \frac{\log(1 + d_{ij}) - \mu_{\text{tr}}}{\sigma_{\text{tr}}}, \\ r_{ij} &= \psi([z_i^{\text{anc}},\, z_{ij},\, z_i^{\text{anc}} - z_{ij},\, z_i^{\text{anc}} \odot z_{ij},\, s_{ij}])\,. \end{aligned} \tag{3}$$

Here d_{ij} is the nonnegative elapsed time in days from prior ECG j to the anchor. The shared multilayer perceptron ψ maps these inputs to relation token r_{ij}. It receives both embed-

dings to retain each ECG's waveform information, their difference to represent change, and their element-wise product to represent interactions between corresponding features. The symbol $\odot$ denotes element-wise multiplication of the two embedding vectors. The scalar s_{ij} is standardized log elapsed time. The log transform limits the influence of very long intervals, and training-fold standardization places time on a dimensionless scale before it enters the relation token, preventing raw values in days from dominating while keeping time continuous. The constants μ_{tr} and σ_{tr} are the training-fold mean and population standard deviation of $\log(1 + d_{ij})$, respectively.

The architecture uses attention-based pooling to assign greater weight to anchor–prior relations that are more informative for prediction while reducing a variable-length ECG history to one patient summary (Ilse et al., 2018). A two-layer scoring network with a hyperbolic-tangent activation assigns each relation token a scalar, and softmax normalization across patient i's prior ECGs produces nonnegative weights α_{ij} that sum to one. The patient-level history summary is

$$u_i = \sum_j \alpha_{ij} r_{ij}. \tag{4}$$

The sum runs over prior ECGs. Exact scorer dimensions and masking details are given in Appendix A. The model's final score ℓ_i, before conversion to a probability, is

$$\ell_i = f_{\text{cur}}(z_i^{\text{anc}}) + \mathbf{1}\{c_i > 1\} f_{\text{hist}}([z_i^{\text{anc}}, u_i]). \tag{5}$$

Here f_{cur} is the anchor-ECG prediction head, f_{hist} is the history-correction head, and $\mathbf{1}\{c_i > 1\}$ is one when a prior ECG is available and zero otherwise. Patients with one eligible ECG use the anchor head alone.

3.3.2. Matched-Pair Loss

Matched-pair loss is the training component of RECAP. ECG count and inter-ECG intervals are shortcut variables that distinguish CS-positive patients from controls without waveform information. Matched-pair loss compares each CS-positive patient with controls having similar ECG count and inter-ECG intervals, making these shortcut variables less useful within each comparison and penalizing the model when the CS-positive patient does not receive a higher score. Here, $\mathcal{P}$ is the set of CS-positive patients, and $\mathcal{M}(i)$ contains up to 100 non-reused controls for patient i, matched on exact ECG count and every aligned inter-ECG interval within seven days (maximum ratio, 100:1) (de Melo Jr et al., 2025). Matched-pair loss adapts the pairwise logistic objective of Burges et al. (2005) to these shortcut-matched case–control sets.

$$\begin{aligned} \mathcal{L}_{\text{MP}} &= \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \frac{1}{|\mathcal{M}(i)|} \sum_{n \in \mathcal{M}(i)} \log\left(1 + \exp\left[-\frac{\ell_i - \ell_n}{T}\right]\right), \\ \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{cls}} + \lambda_{\text{MP}} \mathcal{L}_{\text{MP}}. \end{aligned} \tag{6}$$

Here ℓ_i and ℓ_n are the model scores before conversion to probabilities for CS-positive patient i and matched control n. The pairwise term increases when ℓ_i does not exceed ℓ_n, and the nested averages give each CS-positive patient equal weight. The scalar $T > 0$ is the temperature. The term $\mathcal{L}_{\text{cls}}$ is our study-specific classification loss for the positive-unlabeled

label setting (Bekker and Davis, 2020). It combines focal-modulated binary cross-entropy for patients with documented CS (Lin et al., 2017) with binary cross-entropy that assigns a negative target to patients without documented CS. The scalar λ_{MP} weights matched-pair loss $\mathcal{L}_{\text{MP}}$, and $\mathcal{L}_{\text{total}}$ is the combined objective. Appendix A gives the class-prior weighting, loss hyperparameters, and exact matching and training values.

4. Experimental Setup

We organize the experiments around three questions:

- How predictive are the shortcut variables ECG count and inter-ECG intervals in serial ECG data?
- After controlling for these shortcut variables, do serial ECG waveforms improve prediction over a single ECG?
- How effective are shortcut-aware serial ECG models and training strategies at reducing reliance on these shortcut variables?

4.1. Cohort, Outcome, and ECG Inputs

The source population consists of patients at a tertiary cardiac center who received at least one ECG during January 2019–October 2024. The dataset contains 190 CS-positive patients contributing 936 ECGs and 111,454 patients without documented CS contributing 417,019 ECGs. The resulting cohort reflects routine tertiary cardiovascular care. Patient-level labels indicate whether CS was documented, and only pre-diagnosis ECGs are eligible for CS-positive patients. For patients without documented CS, only ECGs recorded before administrative end of follow-up are eligible. The patient-level outcome is subsequent documented CS diagnosis after the anchor ECG. Each waveform input contains the middle 5000 samples at 500 Hz from the concatenated 12-lead tracing. Detailed waveform processing is described in Appendix A.

4.2. Baseline Designs

4.2.1. Model Benchmarks

To evaluate RECAP, we compare its relational attention architecture with models that differ in their exposure to ECG count and inter-ECG intervals. All serial waveform models use embeddings from the same frozen pretrained ECGFounder checkpoint (Li et al., 2025). Appendix A gives the ECGFounder implementation details, and the benchmark formulations define the interval representations.

Single-ECG models. The Published single-ECG model uses only the anchor waveform and follows the cardiac sarcoidosis model reported by de Melo Jr et al. (2025). The ECGFounder single-ECG model instead applies a trainable linear classification head to the anchor ECG's frozen pretrained ECGFounder representation.

Shortcut-only controls. The Count-only control uses ECG count. The Intervals-only control uses interval count, mean, standard deviation, minimum, maximum, most recent interval, and total span. The Count and intervals control combines ECG count with these inter-ECG-interval summaries.

Shortcut-vulnerable models. These models expose ECG count, inter-ECG intervals, or both through sequence structure or direct interval features.

- **Transformer.** Following the Transformer architecture (Vaswani et al., 2017), the base ECG-only Transformer receives an ordered ECG sequence and adds a learned position embedding that identifies each ECG's position in the sequence. Even without interval inputs, sequence length and its padding mask make ECG count available to the model. To compare interval representations, we also test a Transformer with binned gaps, which maps consecutive inter-ECG intervals to learned bin embeddings, and a Transformer with continuous gaps, which receives the intervals as continuous inputs. Appendix B reports sensitivity to alternative bin widths.
- **Late-fusion ensemble.** Ensemble learning combines predictions from multiple component models (Dietterich, 2000). Our three-branch late-fusion ensemble combines prediction scores from an ECG-only Transformer, a count-and-interval shortcut branch, and the ECGFounder single-ECG model applied to the anchor ECG and held fixed during ensemble training. A trainable linear layer combines the three scores.

Shortcut-aware models. These models aim to reduce shortcut learning by preventing ECG count and inter-ECG intervals from being supplied directly to the prediction head and, when elapsed time is used, incorporating it within waveform-derived representations before pooling.

- **Mean pooling.** ECG mean pooling is a fixed permutation-invariant operator that averages ECG embeddings without sequence-position or interval inputs. Unlike the relational attention architecture, it forms no anchor–prior waveform-relation tokens and uses no learned attention. We also test mean pooling with continuous intervals, which encodes each consecutive inter-ECG interval as a continuous input to the later ECG embedding before averaging all ECG embeddings equally.

4.2.2. Training-Strategy Benchmarks

Each training configuration is evaluated separately, and mitigation strategies are not combined. To evaluate matched-pair loss, we compare it with classification-only training, adversarial shortcut censoring, last-layer retraining, and inter-ECG-interval noise. Each architecture is evaluated under supported standalone configurations. Implementation details are provided in Appendix A for adversarial shortcut censoring, last-layer retraining, and inter-ECG-interval noise.

Adversarial shortcut censoring. Adversarial shortcut censoring (ASC) uses gradient reversal during classification training so that the serial ECG representation becomes less predictive of ECG count and inter-ECG intervals (Ganin et al., 2016; Kim et al., 2019; Jabbour et al., 2020).

Last-layer retraining. Last-layer retraining (LLR) freezes the trained encoder and pooling backbone, reinitializes the prediction head, and optimizes only that head on a separate subset reserved from the training patients, with CS-positive patients and patients without documented CS sampled separately (Kirichenko et al., 2023).

Interval noise. Gaussian jitter adds random noise to time-series inputs during training (Iwana and Uchida, 2021). We adapt it to inter-ECG intervals by adding Gaussian noise to log-transformed intervals during classification training of interval-aware architectures. Interval noise applies only to architectures that consume interval information. Validation and test intervals remain unchanged.

4.3. Training Procedure and Reporting

All train (60%), validation (20%), and test (20%) splits are defined at the patient level. CS-positive patients and patients without documented CS are partitioned separately into five cross-validation folds, and a validation subset is carved from the training portion of each fold. A separate patient subset from the training pool is reserved for last-layer retraining. Within each held-out test fold, patients without documented CS are randomly sampled to target 100 controls per CS-positive patient, forming the unmatched evaluation cohort (de Melo Jr et al., 2025). Reported experiments use Adam with learning rate 3×10^{-4}, weight decay 10^{-5}, and gradient clipping at 1.0. Checkpoints are selected by validation AUROC. The waveform-based models use a study-specific positive-unlabeled training objective, while the shortcut-only controls use class-weighted binary cross-entropy. Matched patient-level AUROC is the primary metric and unmatched AUROC is secondary.

4.4. Shortcut-Controlled Evaluation

The unmatched evaluation shows what each benchmark can recover before shortcut control. The matched evaluation tests whether discrimination remains after CS-positive patients and controls are aligned on ECG count and inter-ECG intervals.

For each held-out CS-positive patient i, eligible controls are defined by

$$\mathcal{M}_\tau(i) = \left\{ n : \begin{array}{l} c_n = c_i, \\ |\delta_{nj} - \delta_{ij}| \le \tau \text{ for all aligned intervals } j = 2, \ldots, c_i, \\ \text{required timestamps are present} \end{array} \right\}. \tag{7}$$

Here i indexes a held-out CS-positive patient, n indexes a candidate control, and $\mathcal{M}_\tau(i)$ is the set of eligible controls. The quantities c_i and c_n are their ECG counts, δ_{ij} and δ_{nj} are their aligned inter-ECG intervals, and τ is the maximum allowed difference in days. Eligible controls have the same ECG count and sufficiently similar intervals. The primary protocol uses exact ECG-count matching and an aligned inter-ECG-interval tolerance of $\tau = 7$ days. Patients with one eligible ECG have no aligned intervals and are matched on ECG count alone. From the unmatched held-out test cohort, controls meeting the matching criteria in Equation 7 were assigned to CS-positive patients without reuse. Balance is assessed on the shortcut variables used to construct the matched cohort. Matched patient-level AUROC is the primary metric, and unmatched patient-level AUROC is secondary. Matched-cohort composition averaged across the three training seeds is reported in Appendix Table 4.

Matching-tolerance sensitivity. Because the primary matched evaluation uses $\tau = 7$ days, we repeat the complete held-out matching and evaluation procedure at $\tau \in \{1, 3, 14, 30\}$ days to assess the stability of our results.

4.5. Shortcut-Aware Component Evaluation

To assess the contribution of each RECAP component, we begin with ECG mean pooling, add anchor–prior relation tokens, incorporate continuous elapsed time, and replace mean pooling with learned attention to form the full model. We also evaluate waveform-only relation-token attention pooling to isolate the contribution of attention. Each architecture is evaluated with classification-only training and matched-pair loss using the same ECG-Founder embeddings, patient-level folds, and evaluation protocol. Appendix A defines the exact architectures.

5. Results

5.1. Shortcut Variables Are Predictive and Inflate Serial ECG Model Performance

ECG count and inter-ECG intervals differed between CS-positive patients and controls prior to matching, indicating the potential for shortcut learning (Figure 2). Consistent with these differences, the Count-only, Intervals-only, and Count and intervals control models achieved AUROCs of 0.760, 0.738, and 0.737, respectively (Table 1). These results show that ECG count and inter-ECG intervals alone carry substantial predictive signal and can provide a shortcut pathway for serial ECG models. After matching, all three controls achieved an AUROC of 0.500, supporting the use of the matched cohort for shortcut-controlled evaluation.

Performance also decreased after matching for every serial ECG model under classification-only training (Table 1). Among shortcut-vulnerable models, matched AUROCs were 0.527–0.619, representing reductions of 0.145–0.206 from unmatched evaluation. Among shortcut-aware models, matched AUROCs were 0.582–0.639, representing reductions of 0.048–0.099. The consistent decline across both model categories provides evidence of shortcut learning through ECG count and inter-ECG intervals.

Matching-tolerance sensitivity. For RECAP, matched AUROC was 0.664–0.672 across alternative tolerances of $\tau \in \{1, 3, 14, 30\}$ days (Appendix Table 3), indicating similar point estimates across the evaluated tolerances. The composition of the primary matched evaluation across three runs is reported in Appendix Table 4.

5.2. RECAP Outperforms Single-ECG Models After Shortcut Control

Table 1 compares the Published single-ECG model and ECGFounder single-ECG model with shortcut-vulnerable and shortcut-aware serial ECG models under unmatched and matched evaluation. RECAP achieves the highest matched AUROC at 0.663 ± 0.051, compared with 0.554 ± 0.015 for the ECGFounder single-ECG model and 0.528 ± 0.045 for the Published single-ECG model.

In the unmatched evaluation, the Published single-ECG model and ECGFounder single-ECG model reach AUROCs of 0.533 and 0.579, respectively. Under classification-only training, the shortcut-vulnerable serial ECG models reach 0.733–0.764, while the shortcut-aware models reach 0.630–0.738. Although these serial models generally outperform the single-ECG models in unmatched evaluation, this difference could be due to shortcut learning.

In the matched evaluation, AUROCs across all serial ECG architectures and training configurations are 0.476–0.663. The relational attention architecture reaches 0.639 under classification-only training, while RECAP achieves the highest matched AUROC of 0.663, compared with 0.554 for the ECGFounder single-ECG model and 0.528 for the Published single-ECG model. This improvement indicates that shortcut-aware modeling of serial ECG waveforms retains predictive information beyond the latest ECG alone.

5.3. RECAP Achieves the Highest Shortcut-Controlled Performance

RECAP, our joint design combining relational ECG attention pooling with matched-pair loss, achieved the highest matched AUROC among the evaluated configurations at 0.663 ± 0.051 (Table 1). Relative to the same architecture under classification-only training, RECAP increased matched AUROC from 0.639 ± 0.044 to 0.663 ± 0.051, while unmatched AUROC decreased from 0.738 ± 0.017 to 0.707 ± 0.030. The strongest alternative was ECG mean pooling with LLR, which reached a matched AUROC of 0.631 ± 0.062.

Across all evaluated training configurations, matched AUROCs were 0.538–0.663 for shortcut-aware models and 0.476–0.619 for shortcut-vulnerable models. Although these ranges overlap, shortcut-aware configurations produced the highest matched results overall. Training-strategy effects varied by architecture, and no strategy improved every supported model over classification-only training. Additionally, bin-width sensitivity for the Transformer with binned gaps produced mixed changes and did not identify a consistently superior binning scheme (Appendix Table 5).

5.3.1. Shortcut-Aware Component Evaluation

When trained with matched-pair loss, matched AUROC increased stepwise as the RECAP components were added (Table 2). Adding anchor–prior waveform-relation tokens to ECG mean pooling increased matched AUROC from 0.588 ± 0.059 to 0.625 ± 0.044. Incorporating continuous elapsed time increased it further to 0.630 ± 0.061, and replacing mean pooling with attention completed RECAP at 0.663 ± 0.051. Waveform-only relation-token attention pooling also improved on waveform-only relation-token mean pooling, reaching 0.638 ± 0.061.

The classification-only results also favored the full architecture used in RECAP, which achieved a matched AUROC of 0.639 ± 0.044. Matched-pair loss produced a clearer incremental pattern, with successive gains from anchor–prior relation modeling, elapsed-time context, and attention pooling.

Table 1: Patient-level AUROC across model architectures and training configurations in the unmatched and shortcut-matched evaluations. Values are mean ± standard deviation across 3 training seeds after averaging across held-out folds. Δ is matched minus unmatched AUROC. Bold marks the highest unmatched and matched AUROC within each model configuration, and underlining marks the highest AUROC overall in each evaluation.

Category	Model family	Model configuration	Training configuration	Unmatched	Matched	Δ (Matched− Unmatched)
Single-ECG	Single-ECG	Published single-ECG model	Classification only	**0.533 ± 0.055**	**0.528 ± 0.045**	−0.005
		ECGFounder single-ECG model	Classification only	**0.579 ± 0.020**	**0.554 ± 0.015**	−0.025
Shortcut-only controls	Controls	Count-only control	Classification only	**0.760 ± 0.006**	**0.500 ± 0.000**	−0.260
		Intervals-only control	Classification only	**0.738 ± 0.006**	**0.500 ± 0.000**	−0.238
		Count + intervals control	Classification only	**0.737 ± 0.002**	**0.500 ± 0.000**	−0.237
Shortcut-vulnerable models	Transformer	ECG-only Transformer	Classification only	**0.739 ± 0.019**	**0.574 ± 0.038**	−0.165
			Matched-pair loss	0.447 ± 0.013	0.559 ± 0.109	+0.112
			ASC	0.710 ± 0.027	0.493 ± 0.070	−0.217
			LLR	0.628 ± 0.047	0.511 ± 0.059	−0.117
		Transformer + binned gaps	Classification only	0.733 ± 0.023	**0.560 ± 0.064**	−0.173
			Matched-pair loss	0.384 ± 0.051	0.524 ± 0.087	+0.140
			ASC	0.739 ± 0.020	0.476 ± 0.067	−0.263
			LLR	0.672 ± 0.028	0.532 ± 0.058	−0.140
			Interval noise	**0.744 ± 0.015**	0.521 ± 0.060	−0.223
		Transformer + continuous gaps	Classification only	**0.733 ± 0.012**	0.527 ± 0.021	−0.206
			Matched-pair loss	0.422 ± 0.071	0.568 ± 0.045	+0.146
			ASC	0.712 ± 0.058	**0.585 ± 0.086**	−0.127
			LLR	0.701 ± 0.045	0.554 ± 0.053	−0.147
			Interval noise	0.722 ± 0.036	0.549 ± 0.054	−0.173
	Late-fusion ensemble	Three-branch late-fusion ensemble	Classification only	**0.764 ± 0.003**	**0.619 ± 0.068**	−0.145
			Matched-pair loss	0.573 ± 0.078	0.579 ± 0.097	+0.006
			ASC	0.762 ± 0.036	0.551 ± 0.127	−0.211
			LLR	0.528 ± 0.049	0.532 ± 0.057	+0.004
			Interval noise	0.764 ± 0.017	0.583 ± 0.096	−0.181
Shortcut-aware models	Mean pooling	ECG mean pooling	Classification only	0.630 ± 0.048	0.582 ± 0.070	−0.048
			Matched-pair loss	0.624 ± 0.013	0.588 ± 0.059	−0.036
			ASC	**0.630 ± 0.039**	0.584 ± 0.070	−0.046
			LLR	0.606 ± 0.034	**0.631 ± 0.062**	+0.025
		Mean pooling + continuous intervals	Classification only	0.661 ± 0.047	0.600 ± 0.033	−0.061
			Matched-pair loss	0.602 ± 0.038	0.579 ± 0.073	−0.023
			ASC	0.622 ± 0.094	0.545 ± 0.072	−0.077
			LLR	0.570 ± 0.060	0.538 ± 0.066	−0.032
			Interval noise	**0.672 ± 0.033**	**0.612 ± 0.031**	−0.060
	Relation-token pooling	Relational ECG attention pooling	Classification only	**0.738 ± 0.017**	0.639 ± 0.044	−0.099
			Matched-pair loss RECAP (ours)	0.707 ± 0.030	**0.663 ± 0.051**	−0.044
			ASC	0.677 ± 0.029	0.562 ± 0.039	−0.115
			LLR	0.713 ± 0.016	0.569 ± 0.030	−0.144
			Interval noise	0.735 ± 0.030	0.617 ± 0.043	−0.118

Table 2: Component evaluation of the relational ECG architecture under classification-only and matched-pair training. Rows add anchor–prior relation tokens, continuous elapsed time, and attention pooling to ECG mean pooling. RECAP denotes the complete relational architecture trained with matched-pair loss. Values are mean ± standard deviation across 3 training seeds, and bold marks the highest AUROC in each column.

Model configuration	Relation tokens	Continuous elapsed time	Attention pooling	Unmatched		Matched	
				Classification only	Matched-pair loss	Classification only	Matched-pair loss
ECG mean pooling	×	×	×	0.630 ± 0.048	0.624 ± 0.013	0.582 ± 0.070	0.588 ± 0.059
Relation-token mean pooling (waveform only)	✓	×	×	0.738 ± 0.029	0.720 ± 0.024	0.629 ± 0.056	0.625 ± 0.044
Relation-token mean pooling (with continuous elapsed time)	✓	✓	×	0.725 ± 0.018	**0.723 ± 0.027**	0.613 ± 0.060	0.630 ± 0.061
Relation-token attention pooling (waveform only)	✓	×	✓	0.737 ± 0.029	0.720 ± 0.031	0.609 ± 0.051	0.638 ± 0.061
Relational ECG attention pooling	✓	✓	✓	**0.738 ± 0.017**	0.707 ± 0.030	**0.639 ± 0.044**	**0.663 ± 0.051**

6. Discussion

This study shows that ECG count and inter-ECG intervals contain substantial outcome-associated signal in serial ECG data. Shortcut-only controls distinguished CS-positive patients from patients without documented CS before matching and yielded chance discrimination after matching, confirming that these shortcut variables were predictive in the unmatched cohort and balanced in the matched cohort. Shortcut-aware serial ECG models retained discrimination after matching, showing that predictive signal from clinical ECG information remained after balancing the targeted shortcut variables.

RECAP produced the highest matched point estimate and higher matched estimates than the single-ECG models, suggesting that serial ECG changes provide information beyond the anchor ECG alone. Across the remaining models, the most effective shortcut-aware training strategy depended on the architecture. Training objectives and temporal representations should therefore be selected jointly and assessed under matched evaluation instead of assuming that one intervention will transfer across models.

The unmatched and matched evaluations provided complementary insights for each model. Unmatched performance reflects discrimination in the observed testing setting, including information from both ECG waveforms and the targeted observation-process variables. Matched performance reflects discrimination among patients with similar ECG counts and aligned inter-ECG intervals. The matched cohort is a targeted stress test rather than a deployment population or an estimate of waveform-only performance. Because matching changes the evaluated population, the difference between the two evaluations combines shortcut control with a population shift. The two evaluations should therefore be interpreted as complementary comparisons rather than decomposed into a proportion of performance attributable to shortcuts.

Inter-ECG intervals can provide clinically meaningful temporal context. Elapsed time helps characterize the rate of waveform change, while testing frequency may reflect clinical deterioration, monitoring intensity, or clinician concern. These variables can also encode institution-specific referral, surveillance, and access patterns that may generalize poorly.

Shortcut-aware modeling should therefore preserve meaningful temporal context while testing whether performance persists after the targeted observation-process differences are balanced.

These findings motivate further study of serial ECG models for opportunistic risk assessment. The endpoint was future documented CS diagnosis rather than independently confirmed disease onset, and matched discrimination remained moderate. Establishing clinical utility will require external and prospective validation, calibration, a prespecified operating point, and evaluation of diagnostic yield, false referrals, clinical workload, net benefit, and subgroup performance. Future clinical studies should test whether model-guided referral, imaging, or treatment decisions improve patient outcomes.

Limitations. This was a retrospective single-center study with a modest number of CS-positive patients and no external validation. Matching retained a smaller and more selected cohort, limiting precision and generalizability. The outcome was future documented CS diagnosis rather than independently confirmed disease onset. Patients without documented CS may therefore include undiagnosed cases, and documented diagnoses may partly reflect referral and diagnostic workup. Matching controlled only ECG count and aligned inter-ECG intervals, leaving other observation-process factors such as referral pathway, care setting, follow-up duration, and clinician concern uncontrolled. The Intervals-only control also partly represents ECG count because the number of intervals depends on the number of ECGs. Variable-length models retain access to history length, and RECAP explicitly represents elapsed time. The matched analysis balances the targeted variables while retaining residual exposure to history length, elapsed time, and other observation-process factors.

Overall, rigorous evaluation of serial ECG models requires explicit shortcut controls and matched evaluation while preserving clinically meaningful temporal information.

Ethics

This retrospective study used data from a single tertiary cardiac center under institutional approval 20240807-01H.

References

Denis Agniel, Isaac S. Kohane, and Griffin M. Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*, 361:k1479, 2018. doi: 10.1136/bmj.k1479.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020. doi: 10.1007/s10994-020-05877-5.

Zeineb Bouzid, Ziad Faramand, Christian Martin-Gill, Susan M. Sereika, Clifton W. Callaway, Samir Saba, Richard Gregg, Fabio Badilini, Ervin Sejdic, and Salah S. Al-Zaiti. Incorporation of serial 12-lead electrocardiogram with machine learning to augment the out-of-hospital diagnosis of non-st elevation acute coronary syndrome. *Annals of Emergency Medicine*, 81(1):57–69, 2023. doi: 10.1016/j.annemergmed.2022.08.005.

Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. Detecting shortcut learning for fair medical ai using shortcut testing. *Nature Communications*, 14(1):4314, 2023. doi: 10.1038/s41467-023-39902-7.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96, 2005. doi: 10.1145/1102351.1102363.

Trenton Chang, Michael W. Sjoding, and Jenna Wiens. Disparate censorship & undertesting: A source of label bias in clinical machine learning. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 343–390, 2022. URL https://proceedings.mlr.press/v182/chang22a.html.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

Richard K. Cheng, Michelle M. Kittleson, Craig J. Beavers, David H. Birnie, Ron Blankstein, Paco E. Bravo, Nisha A. Gilotra, Marc A. Judson, Kristen K. Patton, Leonie Rose-Bovino, and American Heart Association Heart Failure and Transplantation Committee of the Council on Clinical Cardiology, and Council on Cardiovascular and Stroke Nursing. Diagnosis and management of cardiac sarcoidosis: A scientific statement from the american heart association. *Circulation*, 149(21):e1197–e1216, 2024. doi: 10.1161/CIR.0000000000001240.

Ji-Hoon Choi, Sung-Hee Song, Hongryul Kim, Jongwoo Kim, Heesun Park, JaeHu Jeon, JoongSik Hong, Hye Bin Gwag, Sung Ho Lee, Jaichan Lee, Soo Jin Cho, Seung-Jung Park, Young Keun On, Ju Youn Kim, and Kyoung-Min Park. Machine learning algorithm to predict atrial fibrillation using serial 12-lead ecgs based on left atrial remodeling. *Journal of the American Heart Association*, 13(19):e034154, 2024. doi: 10.1161/JAHA.123.034154.

Alessandro De Bortoli, Kristin A. Dawson, Dalia Hashem, Stewart David Spence, Elena Pena, Joao R. Inacio, Pablo Nery, Daniel Juneau, Girish Dwivedi, Robert Beanlands, Ian Paterson, and David H. Birnie. Screening tools for the detection of clinically silent cardiac sarcoidosis. *Respiratory Medicine*, 224:107538, 2024. doi: 10.1016/j.rmed.2024.107538.

Jose F. de Melo Jr, Kathryn E. Mangold, Julia Debertin, Andrew Rosenbaum, John P. Bois, Zachi I. Attia, Paul A. Friedman, Abhishek J. Deshmukh, Suraj Kapa, Leslie T. Cooper, Omar F. Abou Ezzeddine, and Konstantinos C. Siontis. Detection of cardiac sarcoidosis with the artificial intelligence-enhanced electrocardiogram. *Heart Rhythm*, 22 (3):859–861, 2025. doi: 10.1016/j.hrthm.2024.08.013.

Thomas G. Dietterich. Ensemble methods in machine learning. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2000. doi: 10.1007/3-540-45014-9_1.

Jacob Fauber and Christian R. Shelton. Modeling "presentness" of electronic health record data to improve patient state estimation. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 500–513, 2018.

Joseph Futoma, Morgan Simons, Finale Doshi-Velez, and Rishikesan Kamaleswaran. Generalization in clinical prediction models: The blessing and curse of measurement indicator variables. *Critical Care Explorations*, 3(7):e0453, 2021. doi: 10.1097/CCE.0000000000000453.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://jmlr.org/papers/v17/15-239.html.

Richard E. Gregg, Daniel C. Deluca, Cheng-Hao Simon Chien, Eric D. Helfenbein, and Mario Ariet. Automated serial ecg comparison improves computerized interpretation of 12-lead ecg. *Journal of Electrocardiology*, 45(6):561–565, 2012. doi: 10.1016/j.jelectrocard.2012.07.021.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Brian Kenji Iwana and Seiichi Uchida. An empirical survey of data augmentation for time series classification with neural networks. *PLOS ONE*, 16(7):e0254841, 2021. doi: 10.1371/journal.pone.0254841.

Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W. Sjoding, and Jenna Wiens. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *Proceedings of the 5th Machine Learning for Healthcare Conference*, volume 126 of *Proceedings of*

Machine Learning Research, pages 750–782. PMLR, 2020. URL https://proceedings.mlr.press/v126/jabbour20a.html.

Susumu Katsushika, Satoshi Kodera, Mitsuhiko Nakamoto, Kota Ninomiya, Nobutaka Kakuda, Hiroki Shinohara, Ryo Matsuoka, Hirotaka Ieki, Masae Uehara, Yasutomi Higashikuni, Koki Nakanishi, Tomoko Nakao, Norifumi Takeda, Katsuhito Fujiu, Masao Daimon, Jiro Ando, Hiroshi Akazawa, Hiroyuki Morita, and Issei Komuro. Deep learning algorithm to detect cardiac sarcoidosis from echocardiographic movies. *Circulation Journal*, 86(1):87–95, 2022. doi: 10.1253/circj.CJ-21-0265.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Kim_Learning_Not_to_Learn_Training_Deep_Neural_Networks_With_Biased_CVPR_2019_paper.html.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Changxin Lai, Minglang Yin, Eugene G. Kholmovski, Maryam M. Sani, Nisha A. Gilotra, Jonathan Chrispin, and Natalia A. Trayanova. Predicting sudden cardiac death in patients with sarcoidosis using a multimodal artificial intelligence model. *JACC: Clinical Electrophysiology*, 12(3):586–599, 2026. doi: 10.1016/j.jacep.2025.10.009.

Jukka Lehtonen, Valtteri Uusitalo, Pauli Poyhonen, Mikko I. Mayranpaa, and Markku Kupari. Cardiac sarcoidosis: phenotypes, diagnosis, treatment, and prognosis. *European Heart Journal*, 44(17):1495–1510, 2023. doi: 10.1093/eurheartj/ehad067.

Jun Li, Aaron D. Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M. Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7):AIoa2401033, 2025. doi: 10.1056/AIoa2401033.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

Zachary C. Lipton, David Kale, and Randall Wetzel. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. In Finale Doshi-Velez, Jim Fackler, David Kale, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pages 253–270, Northeastern University, Boston, MA, USA, 18–19 Aug 2016. PMLR. URL https://proceedings.mlr.press/v56/Lipton16.html.

Ricards Marcinkevics, Ece Ozkan, and Julia E. Vogt. Debiasing deep chest x-ray classifiers using intra- and post-processing methods. In Zachary Lipton, Rajesh Ranganath,

Mark Sendak, Michael Sjoding, and Serena Yeung, editors, *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 504–536. PMLR, 05–06 Aug 2022. URL https://proceedings.mlr.press/v182/marcinkevics22a.html.

Yu Nishihara, Makoto Nishimori, Satoki Shibata, Masakazu Shinohara, Ken-Ichi Hirata, and Hidekazu Tanaka. Identifying heart failure dynamics using multi-point electrocardiograms and deep learning. *European Heart Journal – Digital Health*, 6(3):447–455, 2025. doi: 10.1093/ehjdh/ztaf016.

Cathy Ong Ly, Balagopal Unnikrishnan, Tony Tadic, Tirth Patel, Joe Duhamel, Sonja Kandel, Yasbanoo Moayedi, Michael Brudno, Andrew Hope, Heather Ross, and Chris McIntosh. Shortcut learning in medical ai hinders generalization: method for estimating ai model generalization without external data. *npj Digital Medicine*, 7(1):124, 2024. doi: 10.1038/s41746-024-01118-4.

Agnese Sbrollini, Marjolein C. De Jongh, C. Cato Ter Haar, Roderick W. Treskes, Sumche Man, Laura Burattini, and Cees A. Swenne. Serial electrocardiography to detect newly emerging or aggravating cardiac pathology: a deep-learning approach. *Biomedical Engineering Online*, 18(1):15, 2019. doi: 10.1186/s12938-019-0630-9.

Agnese Sbrollini, C. Cato ter Haar, Chiara Leoni, Micaela Morettini, Laura Burattini, and Cees A. Swenne. Advanced repeated structuring and learning procedure to detect acute myocardial ischemia in serial 12-lead ecgs. *Physiological Measurement*, 44(8):084003, 2023. doi: 10.1088/1361-6579/ace241.

Rose Sisk, Lijing Lin, Matthew Sperrin, Jessica K. Barrett, Brian Tom, Karla Diaz-Ordaz, Niels Peek, and Glen P. Martin. Informative presence and observation in routine health data: A review of methodology for clinical risk prediction. *Journal of the American Medical Informatics Association*, 28(1):155–166, 2021. doi: 10.1093/jamia/ocaa242.

M. Sunemark, L. Edenbrandt, H. Holst, and L. Sornmo. Serial vcg/ecg analysis using neural networks. *Computers and Biomedical Research*, 31(1):59–69, 1998. doi: 10.1006/cbmr.1997.1462.

Amelia L. M. Tan, Emily J. Getzen, Meghan R. Hutch, Zachary H. Strasser, Alba Gutiérrez-Sacristán, Trang T. Le, Arianna Dagliati, Michele Morris, David A. Hanauer, Bertrand Moal, Clara-Lea Bonzel, William Yuan, Lorenzo Chiudinelli, Priam Das, Harrison G. Zhang, Bruce J. Aronow, Paul Avillach, Gabriel A. Brat, Tianxi Cai, Chuan Hong, William G. La Cava, He Hooi Will Loh, Yuan Luo, Shawn N. Murphy, Kee Yuan Hgiam, Gilbert S. Omenn, Lav P. Patel, Malarkodi Jebathilagam Samayamuthu, Emily R. Shriver, Zahra Shakeri Hossein Abad, Byorn W. L. Tan, Shyam Visweswaran, Xuan Wang, Griffin M. Weber, Zongqi Xia, Bertrand Verdy, The Consortium for Clinical Characterization of COVID-19 by EHR (4CE), Qi Long, Danielle L. Mowery, and John H. Holmes. Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? *Journal of Biomedical Informatics*, 139:104306, 2023. doi: 10.1016/j.jbi.2023.104306.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Nicole G. Weiskopf, Alex Rusanov, and Chunhua Weng. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annual Symposium Proceedings*, 2013:1472–1477, 2013.

Appendix A. Supplemental Methods

Waveform Input Window

Each ECG waveform input contains the middle 5000 samples at 500 Hz from the concatenated 12-lead tracing. For the Published single-ECG model, 120 zeros are appended to obtain the 5120-sample input expected by its implemented architecture (de Melo Jr et al., 2025). This padding serves only to satisfy that fixed input dimension and contains no ECG information.

ECGFounder Embedding Implementation

All waveform-based serial ECG models in Table 1 use the same frozen pretrained ECGFounder checkpoint (Li et al., 2025). Let $e_{ij} \in \mathbb{R}^{1024}$ denote its penultimate representation for ECG j from patient i, and let $e_i^{\text{anc}} = e_{i,c_i}$ denote the corresponding anchor representation. The relational attention architecture applies the same trainable projection block to every ECGFounder representation:

$$z_{ij} = \text{Proj}(e_{ij}), \qquad z_i^{\text{anc}} = \text{Proj}(e_i^{\text{anc}}). \tag{8}$$

The function Proj returns a 256-dimensional task-specific embedding using a 1,024-to-256 linear layer, Gaussian error linear unit activation, dropout of 0.20, and layer normalization. This notation reserves e for frozen ECGFounder representations and z for the projected embeddings used by the relational attention architecture. The projection block and downstream relational attention layers are optimized during serial ECG model training.

Baseline and Benchmark Architecture Formulations

The equations below use the patient index i, chronological ECG index $j = 1, \ldots, c_i$, ECG count c_i, acquisition time t_{ij}, and consecutive interval δ_{ij} defined in the main Methods. The frozen ECGFounder representation is e_{ij}. For continuous consecutive-gap encoding, define

$$\begin{aligned} \delta_{ij}^{\text{enc}} &= \begin{cases} 0, & j = 1, \\ \delta_{ij}, & j = 2, \ldots, c_i, \end{cases} \\ g_{ij} &= \begin{bmatrix} \big(\log(1 + \delta_{ij}^{\text{enc}}) - \mu_\delta\big)/\sigma_\delta \\ \mathbf{1}\{j = 1\} \\ \mathbf{1}\{\text{timing unavailable}\} \end{bmatrix}. \end{aligned} \tag{9}$$

Here g_{ij} is the three-component continuous-gap feature for ECG j, and μ_δ and σ_δ are the training-fold mean and population standard deviation of log consecutive intervals. The first-ECG indicator distinguishes the absence of a preceding interval from an observed zero-day interval. Padding is excluded from all pooling and Transformer calculations. Each architecture maps its final patient representation u_i to a prediction score ℓ_i through a model-specific prediction head.

Single-ECG and shortcut-only controls. Both single-ECG models use only the anchor waveform. The ECGFounder single-ECG model applies ECGFounder to the anchor ECG and trains only its linear classification head, leaving the pretrained encoder fixed. It

is distinct from the separately implemented Published single-ECG model. The Count-only control receives c_i. The Intervals-only control receives interval count, mean, standard deviation, minimum, maximum, most recent interval, and total history span. The Count and intervals control concatenates c_i with these interval summaries.

Mean pooling without and with continuous intervals. ECG mean pooling averages the frozen ECGFounder representations equally:

$$u_i^{\text{mean}} = \frac{1}{c_i} \sum_{j=1}^{c_i} e_{ij}. \tag{10}$$

It uses neither ECG order nor intervals. Mean pooling with continuous intervals first modifies each ECG representation using its preceding continuous-gap feature and then retains equal pooling:

$$e_{ij}^{\text{gap}} = \text{GapFuse}(e_{ij}, g_{ij}), \qquad u_i^{\text{gap-mean}} = \frac{1}{c_i} \sum_{j=1}^{c_i} e_{ij}^{\text{gap}}. \tag{11}$$

The function GapFuse denotes the implemented learned input projection, continuous-gap encoder, residual gate, and layer normalization. Unlike the relational attention architecture, this model forms no anchor–prior relation tokens and learns no pooling weights.

Transformer without and with interval inputs. Let W_{in} denote the ECG input projection and p_j the learned sequence-position embedding for ECG j. Let k_{ij} identify the discrete bin containing δ_{ij}, E_{bin} denote the bin-embedding lookup, and E_{cont} denote the continuous-gap encoder. The three Transformer inputs are

$$h_{ij}^{(0,\text{ECG})} = W_{\text{in}} e_{ij} + p_j, \tag{12}$$

$$h_{ij}^{(0,\text{bin})} = W_{\text{in}} e_{ij} + p_j + E_{\text{bin}}(k_{ij}), \tag{13}$$

$$h_{ij}^{(0,\text{cont})} = W_{\text{in}} e_{ij} + p_j + E_{\text{cont}}(g_{ij}). \tag{14}$$

For each variant, the Transformer processes the ordered input sequence and returns final contextualized ECG representations h_{ij}. Its patient representation is

$$u_i^{\text{Tr}} = \frac{1}{c_i} \sum_{j=1}^{c_i} h_{ij}. \tag{15}$$

The original interval bins use upper boundaries of 1, 7, 30, 90, 180, 365, and 730 days, with separate first-ECG and overflow bins. Appendix Table 5 reports the wider-bin sensitivity analysis.

Late-fusion ensemble. The late-fusion ensemble combines prediction scores from an ECG-only Transformer, a count-and-interval shortcut branch, and the ECGFounder single-ECG model applied to the anchor and held fixed during ensemble training:

$$\ell_i^{\text{ensemble}} = \beta_0 + \beta_{\text{Tr}} \ell_i^{\text{Tr}} + \beta_{\text{shortcut}} \ell_i^{\text{shortcut}} + \beta_{\text{single}} \ell_i^{\text{single}}. \tag{16}$$

The β terms are learned linear-combination parameters. Continuous intervals are not added to this benchmark's Transformer branch.

Adversarial Shortcut Censoring Details

For each patient history, the adversary predicts standardized shortcut targets comprising prior-ECG count and six timing summaries: the most recent log interval, the mean, minimum, maximum, and population standard deviation of consecutive log intervals, and log total history span. A separate binary target indicates missing or invalid timing.

Let $v_{i,\text{count}}$ denote the count target, $v_{i,\text{time}} \in \mathbb{R}^6$ the timing-target vector, and $v_{i,\text{miss}} \in \{0, 1\}$ the missing-time target for patient window i. Hats denote adversary predictions. For training minibatch $\mathcal{B}$, the adversarial loss is

$$\begin{aligned}\mathcal{L}_{\text{ASC}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \Big[& \text{SmoothL1}(\hat{v}_{i,\text{count}}, v_{i,\text{count}}) \\ & + \text{SmoothL1}(\hat{v}_{i,\text{time}}, v_{i,\text{time}}) \\ & + \text{BCELogit}(\hat{v}_{i,\text{miss}}, v_{i,\text{miss}}) \Big].\end{aligned} \tag{17}$$

The vector-valued Smooth L1 term is averaged across its six timing targets. The adversarial-loss weight and gradient-reversal strength are both 1.0. Gradient reversal discourages the serial ECG representation from encoding these shortcut targets while the main classifier is trained for CS prediction.

Last-Layer Retraining Details

Within each cross-validation fold, the split routine allocates 25% of the initial training patients to validation and a separate 25% to LLR. CS-positive patients and patients without documented CS are partitioned separately, and the remaining patients form the stage-one training set. After stage one, we reload the checkpoint with the highest validation AUROC.

For stage two, all parameters outside the final prediction head are frozen. The head is reinitialized using Xavier initialization for linear weights, zero linear biases, unit layer-normalization weights, and zero layer-normalization biases. We optimize only this head for five epochs on the reserved patient subset using Adam with learning rate 3×10^{-4}, weight decay 10^{-5}, and the same positive-unlabeled objective as in stage one. Evaluation uses the stage-two checkpoint with the highest validation AUROC across all serial windows. The LLR subset is drawn entirely from development patients using outcome-stratified sampling rather than the count-and-interval matching used for evaluation.

Inter-ECG-Interval Noise Implementation

Noise is applied online during training only to models that consume interval inputs. Validation and test intervals remain unchanged, and each eligible patient window has probability 0.20 of remaining unmodified. For a modified window,

$$\delta_{ij}^{\text{noise}} = \text{clip}\Big(\exp\Big[\log(1 + \delta_{ij}) + \epsilon_i^{\text{hist}} + \epsilon_{ij}^{\text{int}}\Big] - 1,\, 0,\, U_{\text{train}}\Big), \tag{18}$$

where $\epsilon_i^{\text{hist}} \sim \mathcal{N}(0, 0.55^2)$ is shared across the patient's intervals, $\epsilon_{ij}^{\text{int}} \sim \mathcal{N}(0, 0.30^2)$ is interval-specific, and U_{train} is the training-fold upper clipping limit. The perturbed intervals remain nonnegative.

Relation-Pooling Component-Evaluation Details

Table 2 begins with ECG mean pooling as the no-relation reference. Relation-token mean pooling adds anchor–prior waveform relations, relation-token mean pooling with continuous elapsed time adds elapsed time, and the complete relational attention architecture replaces mean pooling with learned attention. The waveform-only relation-token attention-pooling variant isolates the contribution of attention.

For relation-token mean pooling, prior ECGs receive equal weight:

$$u_i^{\text{rel-mean}} = \begin{cases} \dfrac{1}{c_i - 1} \displaystyle\sum_{j=1}^{c_i-1} r_{ij}, & c_i > 1, \\ 0, & c_i = 1. \end{cases} \tag{19}$$

The relation token r_{ij} is defined in the main Methods and either includes or excludes standardized continuous elapsed time according to the component variant.

For attention pooling, the implemented two-layer scoring network assigns each relation token a scalar score:

$$\text{score}_{ij} = w_2^{\mathsf{T}} \tanh(W_1 r_{ij} + b_1) + b_2, \qquad \alpha_{ij} = \frac{\exp(\text{score}_{ij})}{\sum_{k=1}^{c_i-1} \exp(\text{score}_{ik})}, \qquad u_i = \sum_{j=1}^{c_i-1} \alpha_{ij} r_{ij}. \tag{20}$$

Here $W_1 \in \mathbb{R}^{256\times256}$, $b_1, w_2 \in \mathbb{R}^{256}$, and $b_2 \in \mathbb{R}$ are learned scoring-network parameters. The softmax is evaluated only over prior ECGs; padded entries receive zero weight. Patients without a prior ECG use the anchor prediction alone.

Matched-pair loss implementation. Let $\mathcal{D}_+$ denote training examples from patients with documented CS and $\mathcal{D}_U$ examples from patients without documented CS, which are treated as unlabeled. Let $N_+ = |\mathcal{D}_+|$, $N_U = |\mathcal{D}_U|$, and $\hat{\pi} = N_+/(N_+ + N_U)$. The full-cohort classification loss is

$$\mathcal{L}_{\text{cls}} = \hat{\pi} \frac{1}{N_+} \sum_{i \in \mathcal{D}_+} \text{FocalBCE}(\ell_i, 1) + \frac{1}{N_U} \sum_{i \in \mathcal{D}_U} \text{BCE}(\ell_i, 0). \tag{21}$$

We use focal exponent $\gamma = 1.0$ and unit weight on the unlabeled term. The empirical training-fold factor $\hat{\pi}$ is fixed and is distinct from focal loss's optional class-balancing parameter. Matched-pair loss is applied from epoch 1 through the 50-epoch main training phase with all model parameters trainable. We use $\lambda_{\text{MP}} = 0.75$ and $T = 1.0$, giving $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + 0.75\mathcal{L}_{\text{MP}}$.

Appendix B. Supplemental Results

Matching-Tolerance Sensitivity

The sensitivity analysis evaluates RECAP at alternative aligned inter-ECG-interval tolerances of $\tau \in \{1, 3, 14, 30\}$ days.

Table 3: Sensitivity of RECAP matched AUROC to the tolerance used to align inter-ECG intervals between CS-positive patients and controls. Columns show alternative maximum permitted interval differences. Values are mean ± standard deviation across 3 training seeds.

Model configuration	Training configuration	1 day	3 days	14 days	30 days
RECAP (ours)	Matched-pair loss	0.669 ± 0.036	0.664 ± 0.038	0.672 ± 0.033	0.666 ± 0.032

Matched Evaluation Cohort Composition

The primary matched evaluation used exact ECG-count matching and a seven-day tolerance for every aligned consecutive inter-ECG interval, without control reuse. Table 4 summarizes the matched evaluation cohort composition across three runs.

Table 4: Numbers of CS-positive patients and matched controls in each held-out fold of the shortcut-matched evaluation.

Held-out fold	CS-positive patients, mean ± SD	Matched controls, mean ± SD
1	4.0 ± 2.6	189.0 ± 77.2
2	3.3 ± 2.1	176.3 ± 66.2
3	5.0 ± 0.0	221.0 ± 5.6
4	3.3 ± 1.2	172.3 ± 27.5
5	3.3 ± 1.2	181.7 ± 24.6
Total	**19.0 ± 0.0**	**940.3 ± 41.0**

Transformer Interval-Representation Sensitivity

The original discrete representation uses upper boundaries of 1, 7, 30, 90, 180, 365, and 730 days, with an additional bin for intervals longer than 730 days. The double- and four-times-width variants use boundaries of $\{2, 14, 60, 180, 360, 730, 1460\}$ and $\{4, 28, 120, 360, 720, 1460, 2920\}$ days, respectively.

Table 5: Input-representation ablation for classification-only Transformer models. Rows compare no interval input, continuous intervals, and discretized intervals at three bin widths in the unmatched and shortcut-matched evaluations. Values are mean ± standard deviation across 3 training seeds.

Model family	Model configuration	Training configuration	Unmatched AUROC	Matched AUROC
Transformer	ECG-only Transformer	Classification only	0.739 ± 0.019	0.574 ± 0.038
Transformer	Transformer + binned gaps	Classification only	0.733 ± 0.023	0.560 ± 0.064
Transformer	Transformer + double-width binned gaps	Classification only	0.737 ± 0.025	0.539 ± 0.040
Transformer	Transformer + four-times-width binned gaps	Classification only	0.756 ± 0.018	0.584 ± 0.032
Transformer	Transformer + continuous gaps	Classification only	0.733 ± 0.012	0.527 ± 0.021