biomECG: Learning Transferable ECG Representations from Small Cohorts

Anonymous authors
Paper under double-blind review

Abstract

Self-supervised learning (SSL) has become a standard approach to ECG representation learning, reducing the need for large clinically annotated datasets. Foundation models have been proposed based on very large cohorts. However, they may not be suited to every target population, making lightweight SSL pipelines that work with smaller datasets beneficial. We propose biomECG, an ECG biometric authentication framework for SSL pretraining and downstream evaluation, training compact Siamese encoders to classify pairs of recordings as belonging to the same or different patients. We obtain results comparable to those of a state-of-the-art ECG foundation model developed from over 10 million clinically annotated ECGs, using approximately three orders of magnitude fewer patients for pretraining and over 100-fold fewer parameters. Crucially, frozen encoders combined through late fusion of ECG leads often outperform full fine-tuning at limited and intermediate downstream training sizes. Late fusion also improves mortality discrimination over early fusion, particularly with smaller training cohorts. We evaluate in detail the impact of both the SSL pretraining dataset size and the downstream training dataset size using MIMIC and CODE-15. Pretraining on separate cohorts of similar size yields comparable transfer performance on MIMIC and CODE-15, indicating that representations learned through biometric pretraining capture information relevant across datasets. Code for biomECG, including pretraining and evaluation on MIMIC and CODE-15, is available at `https://github.com/anonymoussubmission1000/biomECG-iclr-anon`.

1 Introduction

The electrocardiogram (ECG) is a cornerstone of cardiovascular diagnosis, yet its interpretation requires expertise (Siontis et al., 2021). Deep learning has improved the automated detection of rhythm and conduction abnormalities (Zhu et al., 2020) and enabled broader cardiac phenotyping from ECGs (Attia et al., 2019). Beyond diagnosis, ECG-based deep learning has also shown promise for prognostic tasks such as mortality prediction (Sau et al., 2024; Lima et al., 2021; Lukyanenko et al., 2025). Owing to the availability of larger ECG datasets, ECG foundation models have been recently proposed, e.g. (Li et al., 2025; Nolin-Lapalme et al., 2025), showing promising results in downstream medical tasks compared to fully supervised approaches (Al-Masud et al., 2026; Wan et al., 2025).

Foundation models are based on self-supervised learning (SSL) approaches in order to limit the need for annotated data. ECG biometrics stands out as a promising SSL approach. Indeed, individual authentication exploits differences in ECG morphology and timing that are sufficiently stable to distinguish patients across recordings (Meltzer & Luengo, 2025; Thiebaud et al., 2026). This has been successfully used for contrastive ECG pretraining (Diamant et al., 2022).

Yet massive scale alone does not guarantee transfer, and parameter scale is not needed for it (Al-Masud et al., 2026). The composition of the pretraining corpus can improve performance in some cohorts while impairing generalization to others, and fine-tuning may not fully recover the loss (Li et al., 2025; Khattak et al., 2026). This thus raises the following question: *Is massive patient scale required to learn informative representations?*

We evaluate the performance of biometry-based SSL for downstream medical tasks inside and across cohorts, with varying sizes of pretraining datasets. As they have been shown to remain strong ECG

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.

encoders for diagnostics (Nonaka & Seita, 2021; Oladipo et al., 2025; Chen, 2025), our models rely on residual networks. We also focus on the impact of lead-wise late fusion, which has shown improved performance over standard 12-lead processing (Aublin et al., 2022; Thiebaud et al., 2026). The obtained representations are evaluated on a mortality prediction task. The impact of data budgets for both pretraining and downstream task, adaptation strategies are assessed, as well as performance across multiple source and target cohorts.

Our main contributions are the following: **(1) State-of-the-art mortality prediction from remarkably small biometric pretraining.** Patient authentication proves to be a highly sample-efficient signal for learning prognostic ECG representations. Across multiple mortality cohorts, our compact encoders perform comparably to a large ECG foundation model despite using approximately *1,000-fold fewer patients for pretraining* and **over 100-fold fewer parameters**. The advantage persists along downstream data budgets and under both frozen transfer and full fine-tuning. **(2) A substantial gain from lead-wise late fusion.** Combining lead-specific representations markedly improves mortality prediction, especially with frozen encoders, and can change which pretrained system performs best. **(3) Assessing performance on transfer across datasets and tasks.** For mortality prediction, biometric prediction shows remarkably consistent results when pretraining and downstream cohorts differ. However, performance on standard age and sex retrieval experiments reveal different behavior, showing that the gains are task dependent.

2 Related Work

2.1 ECG Representation Learning and Foundation Models

Deep learning has improved ECG-based diagnosis and prognostic (Zhu et al., 2020; Attia et al., 2019; Lima et al., 2021). With more ECG datasets becoming available, this has prompted the development of ECG foundation models that aim at learning transferable representations (Li et al., 2025; Nolin-Lapalme et al., 2025; Al-Masud et al., 2026; Coppola et al., 2024; McKeen et al., 2025), using up to 10 million ECGs to train. Performance on clinical tasks of such models were benchmarked, evaluating them on various tasks, using both frozen encoders and fine-tuning (Al-Masud et al., 2026). In particular, it appeared that massive parameter scale may not be necessary for quality of such models for ECGs, and that there may be multiple ways of learning meaningful representations. This raises the question of whether massive *patient* scale is also required.

2.2 ECG-based mortality prediction across populations

Deep-learning based mortality prediction from ECG attracted growing interest since a 2020 seminal paper (Raghunath et al., 2020). A thorough benchmark of such predictors is proposed in Lukyanenko et al. (2025), across MIMIC, CODE-15, and a pediatric cohort, comparing classification with deep survival models and highlighting the value of 1D ResNets and multisite training. Related work linked the difference between ECG-estimated and chronological age to mortality in CODE-15 and external cohorts (Lima et al., 2021). Five-year mortality prediction was evaluated for patient-contrastive based models, and found that transfer depends on pretraining cohort composition (Khattak et al., 2026). In a related but distinct approach, Macierzanka et al. (2025) found that longitudinal dissimilarity between a patient's ECGs added prognostic information on mortality.

2.3 Patient identity as a representation-learning objective

Existing foundation models have been trained using various approaches, from supervised (or weakly supervised) (Li et al., 2025; Liu et al., 2024) to self-supervised (Lou et al., 2023; Na et al., 2024; Kim, 2026), with or without a contrastive approach. Among them, ECG biometrics appears promising. It exploits inter-individual differences in waveform morphology and timing (Melzi et al., 2023). More precisely, identity-related signal features have been shown to persist across recording sessions separated in time, at rest (Ibtehaz et al., 2021; Melzi et al., 2023), and under physiological changes induced by exercise (Thiebaud et al., 2026). Patient authentication is an efficient pretraining signal for clinical tasks: biometrics-based contrastive learning at the patient level, bringing closer recordings from the same person in representation space, provided gains on age, sex, atrial fibrillation, and ventricular hypertrophy prediction (Diamant et al., 2022). However, transfer across cohorts showed varying performance (Khattak et al., 2026).

2.4 LEAD-WISE FUSION AND REPRESENTATION ADAPTATION

In the field of multi-lead ECG biometrics, it clearly appeared that intra-session patient authentication strongly overfitted with respect to cross-session authentication (Melzi et al., 2023). A likely cause was that electrode placement was learned instead of individual characteristics. Lead-wise fusion, termed late fusion, was proposed to overcome this (Thiebaud et al., 2026), following earlier results on ECG abnormality detection (Aublin et al., 2022). The impact of late fusion with respect to early fusion (stacking the leads into a vector time series) has yet to be explored for representation learning.

Adaptation is another important choice: ECG foundation model rankings vary between frozen evaluation and full fine-tuning (Al-Masud et al., 2026). In computer vision, fine-tuning can distort pretrained features and reduce out-of-distribution performance relative to linear probing (Kumar et al., 2022). This motivates evaluating lead-wise fusion alongside both adaptation regimes.

3 BIOMECG: BIOMETRIC PRETRAINING AND TRANSFER

We adapt the Siamese ResNet architecture of Thiebaud et al. (2026) to learn ECG representations through patient authentication and transfer them to downstream prediction. We compare one encoder trained on multiple leads jointly with separate single-lead encoders whose predictions are combined.

3.1 BIOMETRIC PRETRAINING

Let $\boldsymbol{X}_i \in \mathbb{R}^{L\times T}$ denote ECG recording i, comprising L leads and T time samples, and let u_i denote the identity of the corresponding patient. For a selected subset of leads $\mathbb{A}$, we write $\boldsymbol{X}_i^{\mathbb{A}} \in \mathbb{R}^{|\mathbb{A}|\times T}$ for the restricted input. We construct a balanced set $\mathcal{P}$ of pairs of distinct recordings. A pair $(i,j) \in \mathcal{P}$ is positive when both recordings belong to the same patient and negative otherwise, with label $b_{ij} = \mathbf{1}_{\{u_i=u_j\}}$.

For each input lead configuration $\mathbb{A}$, we train a distinct Siamese model whose two branches share the same learnable weights (see Appendix D). Accordingly, $\boldsymbol{\theta}_{\mathbb{A}}$ and $\boldsymbol{\eta}_{\mathbb{A}}$ denote the encoder and projection-module weights specific to configuration $\mathbb{A}$. The ResNet encoder $f_{\boldsymbol{\theta}_{\mathbb{A}}}$ maps each recording to an ECG representation $\boldsymbol{r}_i^{\mathbb{A}}$, and the biometric projection module $p_{\boldsymbol{\eta}_{\mathbb{A}}}$ maps this representation to a biometric embedding $\boldsymbol{z}_i^{\mathbb{A}}$ used only during pretraining. Finally, the pair classifier $g_{\boldsymbol{\psi}_{\mathbb{A}}}$, with weights $\boldsymbol{\psi}_{\mathbb{A}}$, compares the two biometric embeddings and returns a scalar compatibility logit:

$$\boldsymbol{r}_i^{\mathbb{A}} = f_{\boldsymbol{\theta}_{\mathbb{A}}}(\boldsymbol{X}_i^{\mathbb{A}}), \quad \boldsymbol{z}_i^{\mathbb{A}} = p_{\boldsymbol{\eta}_{\mathbb{A}}}(\boldsymbol{r}_i^{\mathbb{A}}), \quad s_{ij}^{\mathbb{A}} = g_{\boldsymbol{\psi}_{\mathbb{A}}}(\boldsymbol{z}_i^{\mathbb{A}}, \boldsymbol{z}_j^{\mathbb{A}}), \quad q_{ij}^{\mathbb{A}} = \sigma(s_{ij}^{\mathbb{A}}). \tag{1}$$

where $q_{ij}^{\mathbb{A}}$ is the predicted probability that the two recordings originate from the same patient. The encoder, projection module, and pair classifier are trained jointly by minimizing binary cross-entropy in logit form,

$$\mathcal{L}_{\text{bio}}^{\mathbb{A}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \ell_{\text{BCE}}(s_{ij}^{\mathbb{A}}, b_{ij}), \qquad \ell_{\text{BCE}}(s,b) = -b \log \sigma(s) - (1-b)\log(1-\sigma(s)). \tag{2}$$

3.2 ENCODER TRANSFER AND ADAPTATION

After biometric pretraining, we discard the projection module and pair classifier and retain only the encoder $\boldsymbol{\theta}_{\mathbb{A}}$. For a given downstream task, a newly initialized MLP head $h_{\boldsymbol{\phi}_{\mathbb{A}}}$, with weights $\boldsymbol{\phi}_{\mathbb{A}}$, receives the encoder representation $\boldsymbol{r}_i^{\mathbb{A}}$, optionally concatenated with task-specific covariates $\boldsymbol{c}_i$:

$$\boldsymbol{r}_i^{\mathbb{A}} = f_{\boldsymbol{\theta}_{\mathbb{A}}}(\boldsymbol{X}_i^{\mathbb{A}}), \quad o_i^{\mathbb{A}} = h_{\boldsymbol{\phi}_{\mathbb{A}}}\left([\boldsymbol{r}_i^{\mathbb{A}}, \boldsymbol{c}_i]\right), \quad \widehat{y}_i^{\mathbb{A}} = \rho(o_i^{\mathbb{A}}). \tag{3}$$

Here, $[\cdot,\cdot]$ denotes feature concatenation and $o_i^{\mathbb{A}}$ is the scalar output of the task head. For binary tasks, $o_i^{\mathbb{A}}$ is a logit and $\rho = \sigma$. For age prediction, the head outputs a standardized age estimate and $\rho(o) = \mu_{\text{age}} + \sigma_{\text{age}} o$, where μ_{age} and σ_{age} are computed from the training cohort. As advocated by the multi-center benchmark of Lukyanenko et al. (2025), mortality prediction additionally uses covariates $\boldsymbol{c}_i = [a_i/100, s_i]$, where a_i is chronological age in years and s_i is the binary sex indicator. Age and sex predictions use the ECG representation alone, without additional covariates.

We compare three adaptation regimes. In *frozen transfer*, the pretrained encoder is kept fixed and in evaluation mode, and only the task head is trained. In *full fine-tuning*, the pretrained encoder

and a newly initialized task head are optimized jointly. The biomECG *scratch* baseline uses the same encoder and head architecture, with both trained from random initialization. This procedure is repeated independently for each downstream task. Further architecture information is provided in Appendix D.

3.3 Early and Prediction-Level Late Fusion

We consider the eight-lead set $\mathbb{L} = \{\mathrm{I}, \mathrm{II}, \mathrm{V1}, \ldots, \mathrm{V6}\}$. The remaining standard limb leads (III, aVR, aVL, and aVF) are linear combinations of leads I and II (Ramirez et al., 2024) and therefore provide no additional independent information. Preliminary experiments further showed no performance benefit from including all 12 leads (see Appendix E.3).

Early fusion. A single Siamese model is pretrained with all eight leads as a joint input ($\mathbb{A} = \mathbb{L}$). The same multilead encoder is then used for downstream adaptation, allowing both pretraining and task learning to exploit cross-lead interactions.

Late fusion. For each lead $\ell \in \mathbb{L}$, we independently pretrain a single-lead Siamese model ($\mathbb{A} = \{\ell\}$) and adapt its encoder with a separate downstream head. Parameters are shared between the two branches within each Siamese model, but not across leads. At inference time, the eight independently obtained predictions are averaged:

$$\widehat{y}_i^{\mathbb{L}} = \frac{1}{|\mathbb{L}|} \sum_{\ell \in \mathbb{L}} \widehat{y}_i^{\{\ell\}}. \tag{4}$$

For binary tasks, this average is taken over post-sigmoid probabilities; for age prediction, it is taken over estimates transformed back to years. This prediction-level fusion has no optimized parameters.

4 Experimental protocol

4.1 Datasets

Table 1: Characteristics of the study cohorts. MIMIC and CODE-15 are used for pretraining and downstream tasks. CPET is used for pretraining only. Percentages use the analyzed patient counts, except mortality, which uses patients with valid survival times (159,072 for MIMIC; 90,240 for CODE-15). IQR: interquartile range.

Characteristic	**MIMIC**	**CODE-15**	**CPET**
Country	US	Brasil	France
Patients analyzed	159,084	90,240	1,359
ECGs / exams	783,247	129,504	1,437
Patients with $\geq$2 ECGs / exams	102,394 (64.4%)	24,572 (27.2%)	64 (4.7%)
Age, years, median [IQR]	58 [43–72]	50 [33–68]	61 [50–70]
Male	75,723 (47.6%)	37,417 (41.5%)	808 (59.5%)
Deaths within 365 days	13,376 (8.4%)	1,156 (1.3%)	–

Table 1 summarizes the cohorts. Biometric pretraining uses MIMIC-IV-ECG (Gow et al., 2023), CODE-15 (Ribeiro et al., 2021), and a private CPET dataset provided by the authors of Thiebaud et al. (2026). CPET (cardiopulmonary exercise testing) assesses cardiorespiratory function through ECG recordings spanning rest, warm-up, progressive exercise, and recovery, providing a complementary setting for biometric learning across changing physiological conditions. As CPET does not include mortality outcomes, it is used for pretraining only. Further cohort, dataset-version, and preprocessing details are provided in Appendix A.

4.2 Model Configurations and Baselines

biomECG configurations. We explore biometric pretraining across source cohorts, pretraining cohort sizes, and input lead configurations. For each source cohort and pretraining budget, we train

a separate biomECG model for each selected lead set $\mathbb{A}$, following the formulation in Section 3. Single-lead models use $\mathbb{A} = \{\ell\}$ for $\ell \in \mathbb{L} = \{\mathrm{I}, \mathrm{II}, \mathrm{V1}, \ldots, \mathrm{V6}\}$; for example, $\mathbb{A} = \{\mathrm{I}\}$, or $\mathbb{A} = \{\mathrm{V1}\}$. Early fusion instead uses the joint eight-lead input $\mathbb{A} = \mathbb{L}$. Late fusion consists of the eight independently pretrained and adapted single-lead models, whose final downstream predictions are averaged, as described in Section 3.3.

Downstream adaptation. For every pretrained configuration, we independently adapt the encoder to each downstream cohort and vary the number of patients available for downstream training. We compare frozen transfer, full fine-tuning, and training from scratch, as defined in Section 3.2. Thus, the experimental design separates the effects of the pretraining source cohort, pretraining budget, lead configuration, downstream cohort, downstream training budget, and adaptation regime.

Comparison with ECGFounder and model size. We compare biomECG with ECGFounder (Li et al., 2025), a state-of-the-art ECG foundation model selected for its strong performance on clinical tasks, including mortality prediction, in an independent benchmark (Al-Masud et al., 2026). ECGFounder was pretrained on more than ten million ECGs with 150 clinical annotation labels. We evaluate ECGFounder's official single-lead and native 12-lead checkpoints and construct an eight-lead late-fusion ensemble over I, II, and V1–V6 to match the biomECG setting. These eight leads retain the independent information in a standard 12-lead ECG, and our preliminary experiments showed no performance benefit from including all 12 leads (see Appendix E.3). For both approaches, frozen adaptation and end-to-end fine-tuning start from the same pretrained checkpoints. A single-lead biomECG contains 211,648 encoder parameters, compared with 30,654,000 for single-lead ECGFounder (approximately $145\times$ more). This comparison assesses practical transfer performance rather than isolating the pretraining objective from differences in model capacity and pretraining data. Further details are reported in Appendix E.

4.3 Downstream Tasks and Training Budgets

Downstream tasks. Our primary task is mortality prediction within 365 days, with target

$$\widetilde{y}_i = \mathbf{1}_{\{E_i=1 \wedge T_i \leq 365\}},$$

where E_i indicates a recorded death and T_i is the recorded time to death in days. See Appendix C for a discussion on censored data. Sex classification and age regression serve as auxiliary tasks. We assess classification performance using AUROC, which measures discrimination across decision thresholds, and age prediction using mean absolute error (MAE) in years. Mortality uses class-weighted binary cross-entropy, sex uses unweighted binary cross-entropy, and age uses mean squared error on targets standardized using the downstream training set. Checkpoints are selected by validation AUROC for classification and validation MAE for age.

biomECG pretraining budget. *Throughout most of this study, biomECG pretraining uses only 1,359 patients.* This budget matches the number of patients available in the CPET dataset. We therefore use the same budget for MIMIC and CODE-15, enabling comparisons between pretraining source cohorts at a fixed patient count. These 1,359 patients comprise both biometric training and validation sets (1,087 and 272 patients, respectively, for the reported seed). This deliberately small pretraining budget contrasts with ECGFounder's pretraining corpus of over ten million ECGs. The only exception is the dedicated pretraining-size experiment in Section 3.1, where we vary the MIMIC pretraining budget. For internal pretraining on MIMIC and CODE-15, all biometric training and validation patients are drawn exclusively from the corresponding downstream training partition (Appendix B).

Downstream training budget. Independently of the biometric pretraining budget, we vary the number of patients available for downstream training. We retain each patient's chronologically first available ECG after preprocessing (Appendix A.3), avoiding over-representation of patients with repeated recordings. For each seed, we create a stratified patient-level training/validation/test split and construct nested, stratified subsets of the downstream training partition (Appendix A.1). *Validation and test sets remain fixed across training subset sizes within each seed*, while partitions vary across seeds. This design evaluates downstream data efficiency while keeping the pretrained encoder initialization fixed for a given pretraining configuration and seed.

5 RESULTS

We first vary the downstream training-set size while keeping biomECG pretraining fixed at only 1,359 patients. We then vary the biometric pretraining cohort size while using the full downstream training set, separating the effects of these two data budgets. Finally, we examine transfer across source cohorts and downstream tasks. Unless otherwise stated, curves show mean test performance across available seeds (see Appendix G.1), and shaded bands indicate ± 1 standard deviation. Full results are presented in Appendix G.3.

5.1 LATE FUSION AND FROZEN TRANSFER SUPPORT STRONG DOWNSTREAM PERFORMANCE

Late Fusion Consistently Outperforms Early Fusion - Fig. 1(a-e) We first compare early and late fusion. As shown in Fig. 1, late fusion improves mortality prediction across models, downstream training-set sizes, and training regimes. Gains are particularly pronounced with small downstream cohorts and generally decrease as the training set grows. For example, with 1,000 training patients and frozen encoders, late fusion increases AUROC from 0.74 to 0.78 for biomECG-1k and from 0.77 to 0.78 for ECGFounder. The overall trend holds for both frozen and fine-tuned encoders. These findings agree with previous results on late fusion (Aublin et al., 2022; Thiebaud et al., 2026).

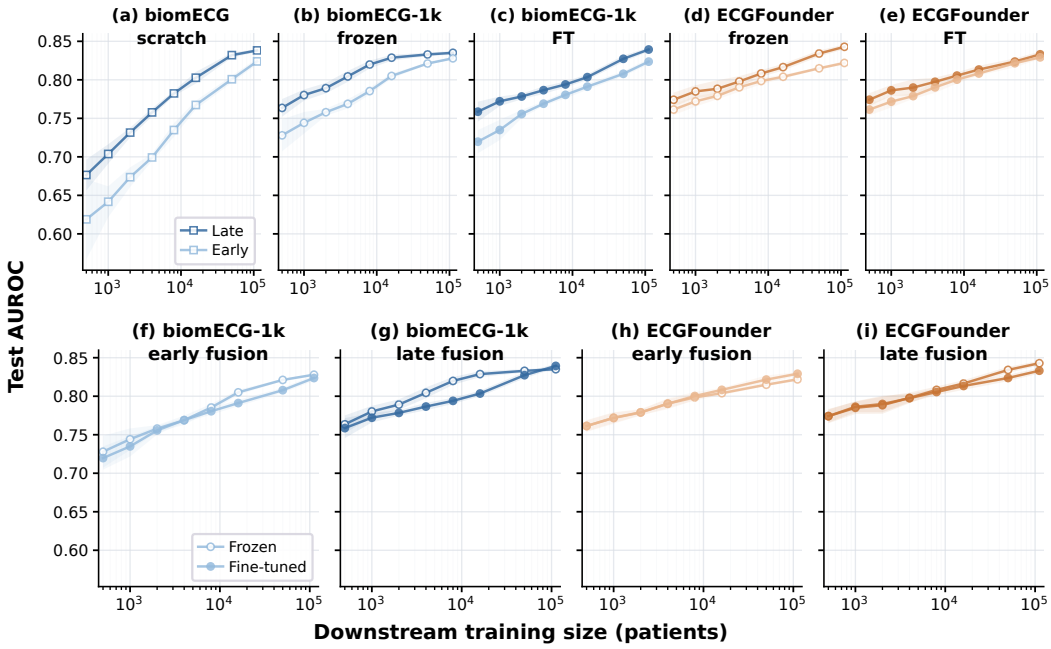

Figure 1: Mortality prediction on MIMIC. Top (a–e): early versus late fusion, by model and downstream training strategy. Bottom (f–i): frozen versus fine-tuned encoders, by model and fusion strategy.

Frozen Transfer Generally Matches or Exceeds Full Fine-Tuning Performance - Fig. 1(f-i) Frozen representations yield mortality prediction performance comparable to or better than fine-tuning across most downstream cohort sizes, particularly for biomECG-1k in the low- and intermediate-data regimes. With late fusion, frozen biomECG-1k reaches 0.80 AUROC with 4,000 training patients, whereas fine-tuning requires 16,000 patients to reach the same performance. Fine-tuning provides only a small improvement at full scale. For ECGFounder, frozen and fine-tuned encoders perform similarly in early fusion, while frozen encoders slightly improve late-fusion AUROC at full scale.

5.2 BIOMETRIC PRETRAINING VS FOUNDATION MODEL

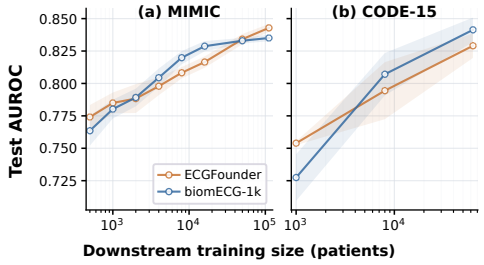

Figure 2: Mortality prediction performance of biomECG-1k and ECGFounder with frozen encoders and eight-lead late fusion on MIMIC and CODE-15.

Comparison with (fixed) small-scale biometric pretraining on two datasets A first major result of this study is that biometric pretraining on only 1,359 patients yields mortality representations competitive with ECGFounder. Fig. 2 compares frozen encoders with late fusion on MIMIC and CODE-15, with biometric pretraining performed on each respective dataset. Prediction heads are trained on nested subsets of increasing size and evaluated on a fixed test set. biomECG-1k outperforms ECGFounder particularly at low and intermediate training sizes on MIMIC and at intermediate and large sizes on CODE-15. For computational reasons, CODE-15 uses two seeds and fewer training sizes, versus five seeds on MIMIC.

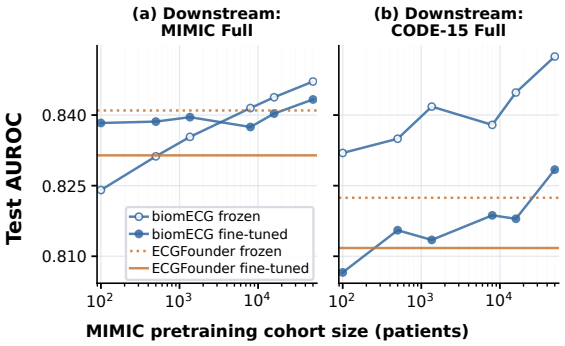

Figure 3: Mortality AUROC on the full downstream training cohorts of (a) MIMIC and (b) CODE-15 versus MIMIC biometric pretraining cohort size. biomECG **is pretrained on MIMIC in both panels**, then frozen or fine-tuned on the downstream dataset with full budget. All methods use eight-lead late fusion and the same single seed (12345) due to limited computational resources.

Varying biometric pretraining cohort size Another major result of the study is that biometric pretraining scales effectively. Having established that pretraining on 1,359 patients is already competitive with ECGFounder, we now vary the MIMIC pretraining cohort from 100 to 50,000 patients to assess how performance depends on scale. We evaluate frozen and fine-tuned encoders on the full downstream mortality training cohorts, using both ECGFounder variants as references. All comparisons use the same single seed for computational reasons. On MIMIC, Fig. 3(a) shows that frozen-encoder AUROC increases steadily from 0.82 with only 100 pretraining patients to 0.85 with 50,000, exceeding both references. Since the encoder is never adapted to mortality, these gains suggest that larger-scale biometric pretraining captures progressively more prognostically relevant information.

An important caveat of the results presented on Fig. 3(a) is that biometric pretraining uses the same cohort later evaluated for mortality, although mortality labels are never exposed. ECGFounder is pretrained on independent cohorts, yet even its downstream fine-tuned variant is outperformed at the largest biometric pretraining scale. To remove this same-cohort advantage, we evaluate the same MIMIC-pretrained encoders on CODE-15, using its full downstream training cohort. Pretraining thus remains exclusively on MIMIC, while mortality training and evaluation use an independent dataset.

In this setting, frozen representations pretrained on just 100 MIMIC patients already achieve 0.83 AUROC, outperforming frozen and fine-tuned ECGFounder (0.82 and 0.81; Fig. 3(b)). Performance improves overall with pretraining scale, reaching 0.85 AUROC with 50,000 patients, and frozen biomECG outperforms its fine-tuned counterpart at every size. Together, these results suggest that biometric pretraining captures substantial prognostically relevant information from very small cohorts and benefits further from scaling. Following this initial cross-dataset experiment, we next examine the influence of the pretraining cohort more systematically.

5.3 TRANSFER OF BIOMETRIC-BASED REPRESENTATIONS ACROSS DATASETS

Fig. 4 evaluates whether biometric representations transfer across pretraining sources. To ensure a fair comparison, all biometric pretraining cohorts contain exactly 1,359 patients, matching the available CPET cohort size. We then evaluate frozen transfer to mortality prediction on both MIMIC and CODE-15 across downstream training-set sizes.

Transfer performance is remarkably consistent across pretraining sources. Biometric pretraining on MIMIC, CODE-15, or CPET yields similar mortality performance on both downstream cohorts; notably, MIMIC pretraining performs best on CODE-15 in several regimes. Strong transfer is also preserved with CPET pretraining, despite ECGs being recorded during graded exercise under widely varying physiological states. These results suggest that biometric pretraining learns cohort-robust features that transfer across datasets and acquisition conditions.

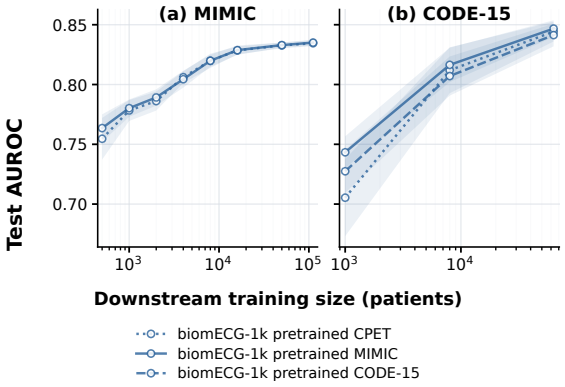

Figure 4: Frozen biomECG-1k mortality prediction with eight-lead late fusion, by biometric pretraining source. Curves and bands show mean test AUROC and one SD on MIMIC (left) and CODE-15 (right; 2 seeds).

5.4 EVALUATION OTHER DOWNSTREAM TASKS: AGE AND SEX PREDICTION

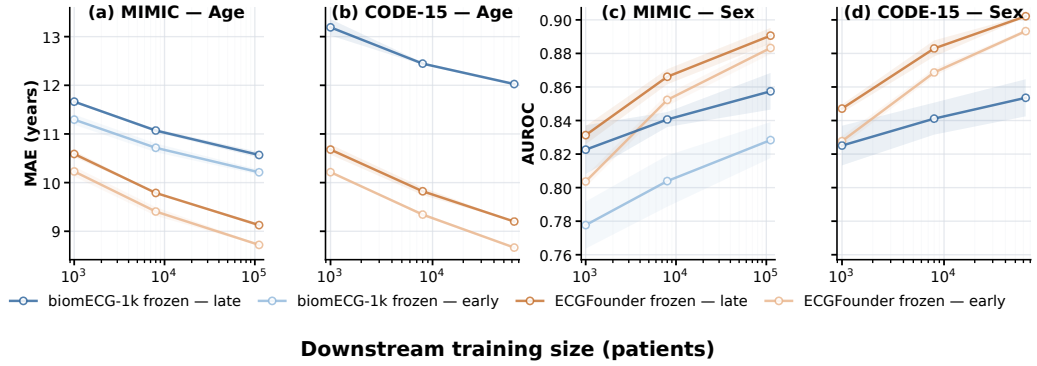

Figure 5: Age (MAE) and sex (AUROC) prediction on MIMIC and CODE-15 using frozen representations. Late fusion averages eight leads; early fusion uses eight for biomECG-1k and twelve for ECGFounder. Curves and bands show mean and one SD; missing results are omitted. 2 seeds.

As a final result, we confront the results observed on mortality to other downstream tasks commonly employed for evaluation: age and sex prediction. Using the same pretrained encoders, we observe markedly different conclusions from mortality. For age prediction, early fusion consistently outperforms late fusion on both MIMIC and CODE-15, including in the low-data regime, and fine-tuning generally improves over frozen representations. ECGFounder also performs better than the biomECG, particularly when the encoders are frozen. A similar pattern is observed for sex prediction: fine-tuning improves performance and ECGFounder remains stronger, especially in the frozen setting. This discrepancy with ranking observed on mortality highlights the task dependence of transfer performance.

6 Discussion and limitations

Patient authentication is a highly data-efficient pretraining signal for prognostic ECG modeling. With only 1,359 pretraining patients, biomECG matches ECGFounder for mortality prediction across downstream regimes and outperforms it in several settings, despite using approximately three orders of magnitude fewer patients and over 100-fold fewer parameters. Scaling biometric pretraining further reaches state-of-the-art-level performance on the full MIMIC cohort. Extending patient-level approaches such as (Diamant et al., 2022), these results show that direct pairwise authentication can yield clinically useful representations from limited pretraining data, without clinical labels or an explicit contrastive loss.

Late fusion and frozen transfer are particularly effective for mortality prediction. Late fusion consistently improves discrimination, especially with limited downstream data, potentially by combining complementary lead-specific experts and reducing prediction variance. The strong performance of frozen encoders further suggests that identity discrimination already captures mortality-relevant information before outcome-specific adaptation, although the underlying ECG characteristics remain to be identified. Training only lightweight prediction heads therefore preserves useful pretrained features while reducing optimization and memory requirements.

Biometric representations improve with scale and transfer across cohorts. Increasing the pretraining cohort improves both biometric performance and frozen mortality prediction, linking stronger identity representations to better prognostic transfer. Comparable transfer from MIMIC, CODE-15, and CPET—including ECGs acquired across rest and exercise—further supports robustness across populations and acquisition conditions.

Transfer behavior is task-dependent. Strategies that performed well for mortality were less effective for age prediction. This suggests that accurate age prediction is not necessarily required for learning mortality-relevant representations. One possible explanation is that the discrepancy between ECG-predicted and chronological age, rather than age-prediction accuracy itself, may carry prognostic information, as suggested by prior work linking this gap to mortality risk (Lima et al., 2021). This hypothesis would require dedicated analysis.

Late fusion increases computation but preserves the lightweight character of biomECG. Eight-lead fusion requires eight independently trained models, yet the complete biomECG ensemble contains only 1.69 million encoder parameters, compared with 245.23 million for late-fusion ECGFounder (Appendix 5). Together with short biometric pretraining, this makes local pretraining and frozen adaptation accessible to clinical teams without foundation-model-scale infrastructure.

Several limitations define the scope of these conclusions. We keep validation sets fixed across nested training cohorts, using them only for checkpoint selection and early stopping to isolate the effect of training-set size; an alternative design could vary a joint train–validation budget. Binary one-year mortality does not model censoring or event times, motivating extensions to survival analysis and cardiovascular-specific outcomes. Clinical evaluation should also include calibration, operating-point utility, and subgroup performance. Generalization beyond the studied populations remains to be established, some experiments use only two seeds, and comparison with one foundation model cannot establish superiority over all architectures or pretraining objectives. We will release code to reproduce our experiments and train biomECG models.

7 Conclusion

(Al-Masud et al., 2026) showed that massive parameter scale was not required for quality ECG representation learning. This works shows that massive patient scale may also not be needed. Small-scale Siamese biometric pretraining produces mortality representations comparable to those of a strong ECG foundation model while using orders of magnitude fewer patients and parameters. Frozen lead-wise experts combined through late fusion are particularly effective when downstream labels are limited, while larger pretraining cohorts provide further gains. Transfer across MIMIC, CODE-15, and CPET supports the robustness of the learned signal across populations and acquisition conditions. The contrasting age and sex results show that fusion and adaptation remain target-dependent, positioning biomECG as a compact and accessible complement to foundation models rather than a universal replacement.

AI USE STATEMENT

In this work, we used generative AI tools to assist with parts of the implementation of downstream experiments and with aspects of literature analysis. We have not used generative AI tools to formulate or refine research hypotheses, adapt the biometric methodology or experiments, develop theoretical or conceptual frameworks, formulate or prove mathematical claims, generate synthetic data, clean or reformat datasets, perform qualitative analysis, or interpret the scientific results.

Additionally, we used generative AI tools to assist with literature search, drafting parts of the manuscript, and editing the text for clarity and readability. The overall research questions, experimental directions, paper structure, scientific arguments, and organization of the results were developed by the authors. AI-assisted code was manually inspected, tested, and checked for consistency with the intended experiments and with the resulting outputs. Literature identified or summarized with AI assistance was subsequently verified manually using conventional scholarly search tools, and the literature review was independently revisited without generative AI assistance to check coverage and accuracy. The manuscript was entirely checked, adapted and rewritten manually. All AI-assisted work was reviewed by the authors. We take responsibility for the final content of this work, including text, claims, code, or other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The authors are not aware of any specific ethical concerns raised by this work beyond those inherent to the use of previously collected clinical ECG data. All datasets were used in accordance with their respective access conditions and governance requirements.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of this work. The main paper provides a detailed description of the proposed method, experimental protocol, datasets, evaluation metrics, and implementation choices. The accompanying open-source code repository, released anonymously on GitHub during the review process[1], includes the complete implementation, dataset download instructions, preprocessing pipelines, model architectures, biomECG pretraining procedure, training and evaluation scripts, and configuration files needed to reproduce our experiments. Additional methodological and experimental details are provided in the appendix and supplementary materials.

ACKNOWLEDGMENTS

This section will be completed after the anonymity period.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2623–2631. ACM, 2019. doi: 10/gf7mzz.

MA Al-Masud, Juan Miguel Lopez Alcaraz, and Nils Strodthoff. Benchmarking ecg fms: A reality check across clinical tasks. In *International Conference on Learning Representations*, volume 2026, pp. 146689–146715, 2026.

Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Brad J Erickson, Suraj Kapa, and Paul A Friedman. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, September 2019. ISSN 0140-6736. doi: https://doi.org/10.1016/s0140-6736(19)31721-0.

[1] https://github.com/anonymoussubmission1000/biomECG-iclr-anon

Pierre G Aublin, Mouin Ben Ammar, Jeremy Fix, Michel Barret, Joachim A Behar, and Julien Oster. Predict alone, decide together: Cardiac abnormality detection based on single lead classifier voting. *Physiological Measurement*, 43(5):054001, 2022. ISSN 1361-6579. doi: 10/hbtxxm.

Yiqiao Chen. A benchmark study of deep learning methods for multi-label pediatric electrocardiogram-based cardiovascular disease classification. *arXiv preprint arXiv:2510.03780*, 2025. doi: https://doi.org/10.48550/arXiv.2510.03780.

Edoardo Coppola, Mattia Savardi, Mauro Massussi, Marianna Adamo, Marco Metra, and Alberto Signoroni. Hubert-ecg as a self-supervised foundation model for broad and scalable cardiac applications. *medRxiv*, pp. 2024–11, 2024.

Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D. Aguirre, Collin M. Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLOS Computational Biology*, 18(2):e1009862, 2022. ISSN 1553-7358. doi: 10/hbtxxc.

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*, September 2023. doi: 10.13026/4nqg-sb35.

Nabil Ibtehaz, Muhammad EH Chowdhury, Amith Khandakar, Serkan Kiranyaz, M Sohel Rahman, Anas Tahir, Yazan Qiblawey, and Tawsifur Rahman. Edith: Ecg biometrics aided by deep learning for reliable individual authentication. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(4):928–940, 2021.

Gul Rukh Khattak, Konstantinos Patlatzoglou, Joseph Barker, Libor Pastika, Boroumand Zeidaabadi, Aidan R. Birdi, Jiayu Huo, Ahmed El-Medany, Hesham Aggour, Yixiu Liang, Antonio H. Ribeiro, Jeffrey Annis, Antonio Luiz Pinho Ribeiro, Junbo Ge, Daniel B. Kramer, Jonathan W. Waks, Evan Brittain, Nicholas Peters, Fu Siong Ng, and Arunashis Sau. Data distribution impacts the performance and generalisability of contrastive learning-based foundation models of electrocardiograms. *PLOS Digital Health*, 5(9):e0001623, September 2026. ISSN 2767-3170. doi: https://doi.org/10.1371/journal.pdig.0001623.

Sehun Kim. Learning general representation of 12-lead ecg with a joint-embedding predictive architecture. *Neural Networks*, pp. 109334, 2026.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

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), June 2025. ISSN 2836-9386. doi: https://doi.org/10.1056/aioa2401033.

Emilly M. Lima, Antônio H. Ribeiro, Gabriela M. M. Paixão, Manoel Horta Ribeiro, Marcelo M. Pinto-Filho, Paulo R. Gomes, Derick M. Oliveira, Ester C. Sabino, Bruce B. Duncan, Luana Giatti, Sandhi M. Barreto, Wagner Meira Jr, Thomas B. Schön, and Antonio Luiz P. Ribeiro. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature Communications*, 12(1), August 2021. ISSN 2041-1723. doi: https://doi.org/10.1038/s41467-021-25351-7.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31949–31963. PMLR, 21–27 Jul 2024.

Yu-Sheng Lou, Chin-Sheng Lin, Wen-Hui Fang, Chia-Cheng Lee, and Chin Lin. Extensive deep learning model to enhance electrocardiogram application via latent cardiovascular feature extraction from identity identification. *Computer Methods and Programs in Biomedicine*, 231:107359, April 2023. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2023.107359.

Platon Lukyanenko, Joshua Mayourian, Mingxuan Liu, John K. Triedman, Sunil J. Ghelani, and William G. La Cava. Deep survival analysis from adult and pediatric electrocardiograms: a multi-center benchmark study. *BioData Mining*, 19(1), December 2025. ISSN 1756-0381. doi: https://doi.org/10.1186/s13040-025-00510-4.

Krzysztof Macierzanka, Arunashis Sau, Konstantinos Patlatzoglou, Libor Pastika, Ewa Sieliwonczyk, Mehak Gurnani, Nicholas S Peters, Jonathan W Waks, Daniel B Kramer, and Fu Siong Ng. Siamese neural network-enhanced electrocardiography can re-identify anonymized healthcare data. *European Heart Journal - Digital Health*, 6(3):417–426, 2025. ISSN 2634-3916. doi: 10/hbtxwj.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *JAMIA open*, 8(5):ooaf122, 2025.

David Meltzer and David Luengo. Ecg-based biometric recognition: A survey of methods and databases. *Sensors*, 25(6):1864, 2025. ISSN 1424-8220. doi: 10/hbtxwb.

Pietro Melzi, Ruben Tolosana, and Ruben Vera-Rodriguez. Ecg biometric recognition: Review, system proposal, and benchmark evaluation. *IEEE Access*, 11:15555–15566, 2023. ISSN 2169-3536. doi: 10/hbtxv8.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International conference on learning representations*, volume 2024, pp. 15012–15035, 2024.

Alexis Nolin-Lapalme, Achille Sowa, Jacques Delfrate, Olivier Tastet, Denis Corbin, Merve Kulbay, Derman Ozdemir, Marie-Jeanne Noël, François-Christophe Marois-Blanchet, François Harvey, Surbhi Sharma, Minhaj Ansari, I-Min Chiu, Valentina Dsouza, Sam F. Friedman, Michaël Chassé, Brian J. Potter, Jonathan Afilalo, Pierre Adil Elias, Gilbert Jabbour, Mourad Bahani, Marie-Pierre Dubé, Patrick M. Boyle, Neal A. Chatterjee, Joshua Barrios, Geoffrey H. Tison, David Ouyang, Mahnaz Maddah, Shaan Khurshid, Julia Cadrin-Tourigny, Rafik Tadros, Julie Hussin, and Robert Avram. Foundation models for generalizable electrocardiogram interpretation: comparison of supervised and self-supervised electrocardiogram foundation models. *medRxiv*, March 2025. doi: https://doi.org/10.1101/2025.03.02.25322575.

Naoki Nonaka and Jun Seita. In-depth benchmarking of deep neural network architectures for ecg diagnosis. In Ken Jung, Serena Yeung, Mark Sendak, Michael Sjoding, and Rajesh Ranganath (eds.), *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pp. 414–439. PMLR, 2021. URL https://proceedings.mlr.press/v149/nonaka21a.html.

Eyiara Oladipo, Sarwar Nazrul, and Mohamed Nafea. Benchmarking deep learning architectures for ecg-based multi-label heart disease prediction using mimic-iv database. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 795–800. IEEE, 2025. doi: 10/hbtxwp.

Sushravya Raghunath, Alvaro E. Ulloa Cerna, Linyuan Jing, David P. vanMaanen, Joshua Stough, Dustin N. Hartzel, Joseph B. Leader, H. Lester Kirchner, Martin C. Stumpe, Ashraf Hafez, Arun Nemani, Tanner Carbonati, Kipp W. Johnson, Katelyn Young, Christopher W. Good, John M. Pfeifer, Aalpen A. Patel, Brian P. Delisle, Amro Alsaid, Dominik Beer, Christopher M. Haggerty, and Brandon K. Fornwalt. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine*, 26(6):886–891, May 2020. ISSN 1546-170X. doi: https://doi.org/10.1038/s41591-020-0870-z.

Elisa Ramirez, Samuel Ruiperez-Campillo, Ruben Casado-Arroyo, José Luis Merino, Julia E. Vogt, Francisco Castells, and José Millet. The art of selecting the ecg input in neural networks to classify heart diseases: A dual focus on maximizing information and reducing redundancy. *Frontiers in Physiology*, 15, 2024. ISSN 1664-042X. doi: 10/hbtxxb.

Antônio H. Ribeiro, Gabriela M.M. Paixao, Emilly M. Lima, Manoel Horta Ribeiro, Marcelo M. Pinto Filho, Paulo R. Gomes, Derick M. Oliveira, Wagner Meira Jr, Thömas B Schon, and Antonio Luiz P. Ribeiro. Code-15%: a large scale annotated dataset of 12-lead ecgs. June 2021. doi: 10.5281/zenodo.4916206. URL https://doi.org/10.5281/zenodo.4916206.

Arunashis Sau, Libor Pastika, Ewa Sieliwonczyk, Konstantinos Patlatzoglou, Antônio H Ribeiro, Kathryn A McGurk, Boroumand Zeidaabadi, Henry Zhang, Krzysztof Macierzanka, Danilo Mandic, Ester Sabino, Luana Giatti, Sandhi M Barreto, Lidyane do Valle Camelo, Ioanna Tzoulaki, Declan P O'Regan, Nicholas S Peters, James S Ware, Antonio Luiz P Ribeiro, Daniel B Kramer, Jonathan W Waks, and Fu Siong Ng. Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: A model development and validation study. *The Lancet Digital Health*, 6(11):e791–e802, 2024. ISSN 2589-7500. doi: 10/g8sd6s.

Konstantinos C. Siontis, Peter A. Noseworthy, Zachi I. Attia, and Paul A. Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7):465–478, February 2021. ISSN 1759-5010. doi: https://doi.org/10.1038/s41569-020-00503-2.

Luca Thiebaud, Paul Chauchat, Mustapha Ouladsine, and Stéphane Delliaux. Learning cardiac features: Ecg biometrics across time and exercise. In *International Conference on AI in Healthcare*, pp. 94–107. Springer, 2026.

Zhijiang Wan, Qianhao Yu, Jia Mao, Wenfeng Duan, and Cheng Ding. Openecg: Benchmarking ecg foundation models with public 1.2 million records, 2025.

Hongling Zhu, Cheng Cheng, Hang Yin, Xingyi Li, Ping Zuo, Jia Ding, Fan Lin, Jingyi Wang, Beitong Zhou, Yonge Li, Shouxing Hu, Yulong Xiong, Binran Wang, Guohua Wan, Xiaoyun Yang, and Ye Yuan. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. *The Lancet Digital Health*, 2(7):e348–e357, July 2020. ISSN 2589-7500. doi: https://doi.org/10.1016/s2589-7500(20)30107-2.

A DATASETS

A.1 COHORT CHARACTERISTICS

Table 2: Cohort characteristics. CODE-15 is an analyzed subset of its processed source. Patient characteristics use analyzed patients; survival statistics use patients with valid survival times (MIMIC: 159,072; CODE-15: 90,240). The historical MIMIC positive label includes 12 negative survival times. IQR: interquartile range; –: not applicable.

Characteristic	**MIMIC**	**CODE-15**	**CPET**
Patients in processed source	159,084	94,613	1,359
ECGs / exams in processed source	783,247	136,760	1,437
Patients analyzed	159,084	90,240	1,359
ECGs / exams analyzed	783,247	129,504	1,437
Patients with $\geq$2 ECGs / exams	102,394 (64.4%)	24,572 (27.2%)	64 (4.7%)
ECGs / exams per patient, median [IQR]	2 [1–5]	1 [1–2]	1 [1–1]
Age, years, median [IQR]	58 [43–72]	50 [33–68]	61 [50–70]
Male	75,723 (47.6%)	37,417 (41.5%)	808 (59.5%)
Negative survival time	12 (0.0%)	0 (0.0%)	–
Deaths during available follow-up	25,087 (15.8%)	3,161 (3.5%)	–
Deaths within 365 days	13,376 (8.4%)	1,156 (1.3%)	–
Historical positive label (TTE $\leq$365 days)	13,388 (8.4%)	1,156 (1.3%)	–
Censored before 365 days	28,375 (17.8%)	3,878 (4.3%)	–
Time to death, days, median [IQR]	317 [57–1255]	551 [216–959]	–
Time to death within 365 days, median [IQR]	66.5 [15–181]	142 [53–250]	–
Time to censoring, days, median [IQR]	380 [365–1428]	1172 [746–1617.5]	–
Survival time clipped to one day	–	34 (0.0%)	–

Table 3: Unique patients in the saved downstream partitions. Splits are repetead multiple times across seeds.

Cohort	**Train (full)**	**Validation**	**Test**	**Total**
MIMIC	111,358	23,863	23,863	159,084
CODE-15	63,168	13,536	13,536	90,240

A.2 SIGNAL PREPROCESSING.

We harmonized MIMIC-ECG and CODE-15 recordings to 500 Hz and a fixed length of 4,096 samples (8.192 seconds). MIMIC recordings retain their native sampling rate; CODE-15 recordings are resampled from 400 to 500 Hz using polyphase resampling after removing leading and trailing zero padding. Recordings with insufficient signal duration, including seven-second CODE-15 ECGs, are excluded. Each lead is cleaned independently using NeuroKit2's `ecg_clean` function with `method="neurokit"`, applying a fifth-order Butterworth high-pass filter at 0.5 Hz followed by the default 50 Hz powerline filter. We retain the central 4,096 samples and store the signals in single precision. Before model input, each lead is Z-scored independently. The same preprocessing and normalization are used for biomECG and ECGFounder models.

A.3 CODE-15 AND MIMIC-ECG PREPROCESSING

We used the publicly available mortality-prediction implementation associated with Lukyanenko et al. (2025), pinned to commit `199eab6` of the `cavalab/ecg-survival-benchmark` repository. Following the repository documentation, data preparation used MIMIC-IV v2.2 together with MIMIC-IV-ECG v1.0, while CODE-15 was obtained from Zenodo record 4916206.

B MORTALITY PREDICTION: PARTITIONS AND TRAINING BUDGETS

After selecting one ECG per patient, MIMIC contains 159,084 patients and CODE-15 contains 90,240. For each seed, the patient-level train/validation/test splits contain 111,358/23,863/23,863 patients for MIMIC and 63,168/13,536/13,536 for CODE-15. Within each training split, nested subsets are stratified to preserve the mortality prevalence of the full training set, up to integer rounding. Validation and test sets remain fixed across downstream training-set sizes within a seed.

On MIMIC, the downstream mortality training sizes are 500, 1,000, 2,000, 4,000, 8,000, 16,000, 50,000, and full; age and sex use 500, 1,000, 2,000, 8,000, 50,000, and full. To limit computational cost, CODE-15 uses 1,000, 8,000, and full training patients for all three downstream tasks.

For internal biometric pretraining in either cohort, eligible patients are drawn exclusively from the corresponding downstream training partition and must have at least two ECG sessions. A pretraining budget of n comprises the first n eligible patients in the nested training order; no downstream validation or test patient is used. The reference budget contains 1,359 patients, matching CPET, and is divided into 1,087 pretraining and 272 validation patients for the reported seed. On MIMIC, we further evaluate biometric pretraining budgets of 100, 500, 1,359, 8,000, 16,000, and 50,000 patients.

C SHORT FOLLOW-UP AND BINARY LABELS

To align with the binary-label convention of the ECG survival benchmark (Lukyanenko et al., 2025), patients without a recorded death within 365 days are assigned to the negative class, including those censored before 365 days. This concerns 28,375 of the 159,072 MIMIC patients with valid survival times (17.8%) and 3,878 of 90,240 CODE-15 patients (4.3%). Inspection of the data indicates that most patients censored before one year have a single recorded ECG and a time to event of zero days. Treating these patients as event-free for one year may therefore be a reasonable practical approximation for many cases, but its effect on performance was not evaluated and some unobserved deaths may be mislabeled.

D BIOMECG MODEL ARCHITECTURE

D.1 BIOMETRIC PRE-TRAINING

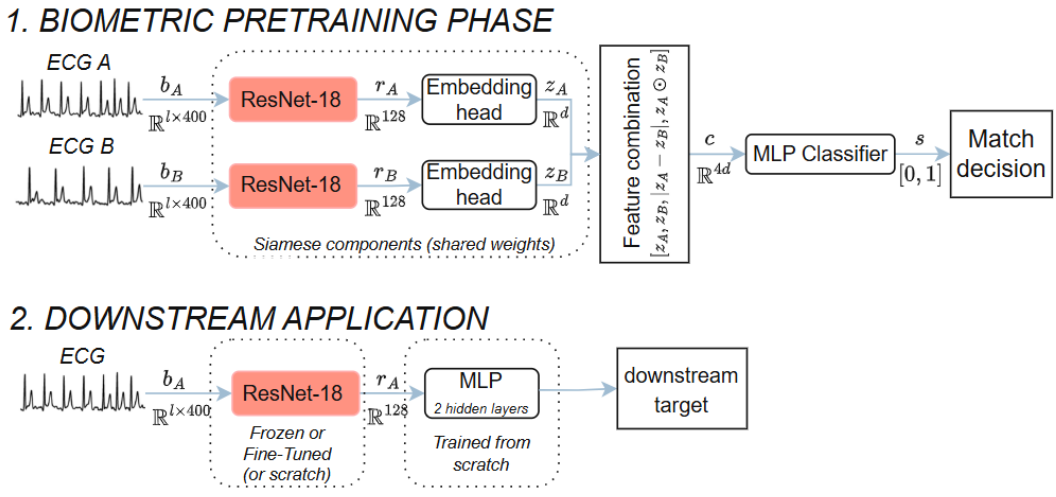

Figure 6: biomECG workflow. Biometric pretraining adapted from Thiebaud et al. (2026), and encoder used for downstream task. Here $l = |\mathbb{A}|$.

D.1.1 ARCHITECTURE

We use a Siamese verification architecture described on Figure 6 adapted from Thiebaud et al. (2026). Two 8.192-s ECG recordings $x_A, x_B \in \mathbb{R}^{|\mathbb{A}|\times 4096}$, sampled at 500 Hz, are processed by a shared 1D ResNet-18 with fixed feature dimension: all four residual stages retain 64 channels, rather than progressively increasing their width. Concatenated adaptive max and average pooling produces 128-dimensional representations. A projection head comprising a linear layer, batch normalization, dropout, and ReLU maps these representations to d-dimensional embeddings. The concatenation $[z_A, z_B, |z_A - z_B|, z_A \odot z_B]$ is passed to a verification MLP with one 256-unit hidden layer and a scalar output. A sigmoid converts its output into the probability that the recordings originate from the same individual.

Biometric pretraining was capped at 15 epochs with early stopping (patience 7). Learning rate, batch size, dropout, convolutional kernel size k, and embedding dimension d were optimized with Optuna (Akiba et al., 2019) using five-fold cross-validation in a preliminary study on lead-II recordings from the CPET training cohort. The selected configuration was shared across single-lead models; the multi-lead early-fusion model was optimized independently. Hyperparameters are presented in Table 4.

D.1.2 ECG PAIRS CONSTITUTION

The Siamese network is trained on pairs of ECG segments: positive pairs originate from the same patient, whereas negative pairs originate from different patients. Patients are partitioned before pair construction to prevent identity overlap between training, validation, and test sets. Positive pairs preferentially combine recordings from different sessions; within-session pairing is used when only one session is available.

For MIMIC and CODE-15 biometric pretraining, we use patients with multiple ECG recordings and construct positive pairs exclusively across recordings. For CPET, a session corresponds to an exercise test containing multiple 10-s strips acquired at different exercise intensities. Positive pairs span different tests when available; otherwise, they combine distinct strips from the same test. Each network input is an 8.192-s segment derived from a strip or ECG recording. Thus, patients define identities, sessions and recordings define acquisition occasions, strips provide signal segments, and pairs constitute the verification training examples.

D.2 DOWNSTREAM ARCHITECTURE

For downstream prediction (see Tab. 6), we discard the projection head and verification classifier and retain the 128-dimensional pooled backbone features. Each task uses a three-layer MLP with two hidden layers of 256 and 128 units, ReLU activations, and a scalar output, without dropout or batch normalization. For mortality, normalized age and binary sex are concatenated with the ECG features, yielding a 130-dimensional input; age and sex prediction use ECG features alone. Binary tasks apply a sigmoid to the output, whereas age prediction uses the scalar directly. We either freeze the backbone, including its batch-normalization statistics, or fine-tune it jointly with the prediction head. The scratch baseline trains both components from random initialization. Early fusion jointly encodes eight leads; late fusion averages the predictions of eight independently trained single-lead models.

Model size. Table 5 reports parameter counts computed directly from the implementations used in our experiments, and for ECGFounder (described in Section E). The latter counts were additionally verified against the local checkpoint tensors. The comparison therefore places compact biometric encoders against a substantially larger model developed with extensive ECG-specific clinical supervision.

Table 4: Biometric pretraining and downstream hyperparameters. Pretraining values apply to both single-lead and early-fusion models. Downstream heads have two hidden layers. Frozen encoders are not updated.

Hyperparameter	**Biometric pretraining**	**Downstream**
Backbone channels per residual stage	64, 64, 64, 64	64, 64, 64, 64
Convolutional kernel size k	3	3
Pooled feature dimension	128	128
Projection dimension d	256	–
MLP hidden widths	256	256, 128
Projection / verification dropout	0.11	–
Batch size	256	256
Pretraining learning rate	2.9×10^{-4}	–
Encoder learning rate (FT / scratch)	–	10^{-4}
Prediction-head learning rate	–	10^{-3}
Maximum epochs	15	200
Early-stopping patience	7	20

Table 5: Encoder parameter counts. Representation dimensions are reported per encoder. Counts include registered encoder parameters and exclude downstream prediction heads, the original ECG-Founder classification layer, and the biometric projection and pair-classification heads. Ensemble counts sum the parameters of eight separate encoders. Frozen encoders retain these parameters but do not optimize them. These counts describe model capacity, rather than computational cost.

Model configuration	Encoders	Representation dimension	Encoder parameters
biomECG, single lead	1	128	211,648
biomECG, eight-lead early fusion	1	128	212,992
biomECG, eight-lead late fusion	8	128	1,693,184
ECGFounder, single lead	1	1,024	30,654,000
ECGFounder, 12-lead early fusion	1	1,024	30,665,264
ECGFounder, eight-lead late fusion	8	1,024	245,232,000

E ECGFounder

E.1 Sources, architecture, and adaptation

Pretraining data and supervision. ECGFounder was developed from the Harvard–Emory ECG Database (HEEDB), comprising 10,771,552 ECGs from 1,818,247 individuals. The authors report a development set of 7,519,035 ECGs from 1,319,128 patients after preprocessing and exclusions, and a separate test set of 834,926 ECGs. We therefore distinguish the size of the original database from the size of the reported development set. Pretraining uses 150 labels extracted from clinical ECG annotations, covering diagnoses, rhythms, and morphological findings. MIMIC-IV-ECG is described as a downstream fine-tuning and evaluation dataset, rather than a source of pretraining data; our experiments initialize from the released pretrained weights (Li et al., 2025).

Checkpoint provenance. We use the two official checkpoints distributed by the authors:

- Single-lead ECGFounder: `1_lead_ECGFounder.pth`.
- 12-lead ECGFounder: `12_lead_ECGFounder.pth`.

The implementation is taken from the official ECGFounder repository, at commit `04edac702b61c91face519774ddcc0cd712fef23`. Both adaptation regimes use the same source weights.

Lead configurations. We distinguish three configurations:

- **ECGFounder-12L:** the official 12-lead checkpoint processes all twelve leads jointly, allowing interactions between leads within the encoder.
- **ECGFounder-1L:** the official single-lead checkpoint initializes a separate downstream model for each evaluated lead.
- **ECGFounder-8L:** eight independently adapted single-lead models process I, II, and V1–V6. Their predictions are combined by an unweighted arithmetic mean. For classification, we average predicted probabilities; for age regression, we average predicted ages. This configuration is our late-fusion construction, rather than a separately released eight-lead checkpoint.

Each expert in the eight-lead ensemble starts from the same single-lead checkpoint, but has its own downstream head and, when fine-tuned, independently updated encoder weights. Fusion introduces no additional trainable parameters. The eight-lead comparison matches the input leads across model families; the native ECGFounder-12L configuration additionally uses III, aVR, aVL, and aVF.

Model size. Table 5 reports parameter counts for ECGFounder, which were additionally verified against the local checkpoint tensors.

E.2 ECGFounder downstream adaptation

We use the pooled ECGFounder representation in place of its original 150-output prediction head. A task-specific multilayer perceptron follows the same hidden-layer structure as for biomECG: $D_{\text{in}} \rightarrow 256 \rightarrow 128 \rightarrow 1$, with ReLU activations between linear layers. Mortality models concatenate age and sex to the ECG representation; age and sex prediction models use ECG features alone. Consequently, the head topology is shared, but its parameter count differs because the representation dimensions differ. Frozen adaptation trains only this head while keeping the encoder in evaluation mode; fine-tuning also updates the encoder.

Both model families receive the same preprocessed ECG windows of 4,096 samples at 500 Hz, corresponding to 8.192 seconds, with per-recording, per-lead standardization. This keeps the signal duration and preprocessing consistent across the comparison

E.3 Eight versus Twelve Leads

We use the eight independent leads (I, II, and V1–V6), since the remaining limb leads (III, aVR, aVL, and aVF) are linear combinations of leads I and II. Preliminary analysis confirms equivalent performance between the eight- and twelve-lead ECGFounder configurations. This enables a direct comparison between late-fusion ECGFounder and biometric ECG models in the same eight-lead setting.

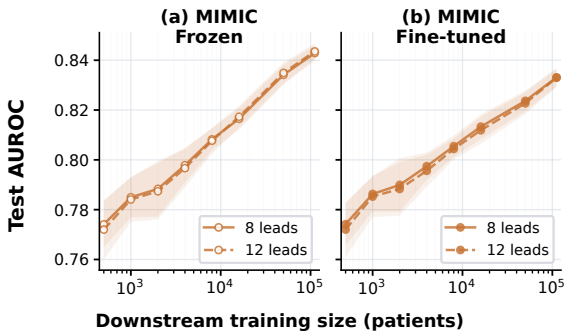

Figure 7: ECGFounder mortality prediction on MIMIC with frozen and fine-tuned encoders. Late fusion averages either eight experts (I, II, V1–V6; solid) or all twelve (dashed). Curves and bands show mean test AUROC and one SD using matched runs at each training size.

F BIOMETRIC PERFORMANCES

F.1 EQUAL ERRORS RATES OF PRETRAINED BIOMETRIC ARCHITECTURES

Table 6: Biometric verification on cross-session test pairs: EER (%), mean (SD) across available seeds. The equal error rate (EER) is the operating point where false acceptance and false rejection rates are equal; lower is better. Each column denotes a separately trained encoder; Early (8L) jointly encodes eight leads. Increasing the MIMIC pretraining cohort generally reduces EER, indicating a scaling benefit. Two seeds: CODE-15 1,359, MIMIC 100, MIMIC 500, MIMIC 8,000, MIMIC 16,000, MIMIC 50,000.

Model	Dataset	Pretraining patients	I	II	V1	V2	V3	V4	V5	V6	Early (8L)
biomECG	MIMIC	100	21.8 (0.083)	22.5 (1.5)	23.6 (2.1)	27.5 (1.3)	29.0 (0.32)	26.3 (1.2)	25.3 (2.1)	28.1 (0.85)	24.4 (2.0)
biomECG	MIMIC	500	14.0 (1.2)	14.2 (2.7)	16.8 (1.4)	20.7 (2.4)	20.2 (2.5)	17.8 (2.3)	17.0 (0.86)	19.3 (3.2)	12.4 (0.98)
biomECG	MIMIC	1,359	11.9 (0.90)	12.3 (1.0)	14.9 (0.85)	16.4 (0.79)	16.6 (0.87)	14.7 (0.62)	14.5 (0.70)	15.1 (0.89)	9.98 (0.63)
biomECG	MIMIC	8,000	8.06 (0.18)	9.13 (0.087)	10.9 (0.70)	11.4 (1.3)	11.8 (1.2)	10.0 (0.73)	10.3 (0.33)	10.4 (0.92)	5.14 (0.56)
biomECG	MIMIC	16,000	8.13 (0.92)	7.69 (0.61)	9.62 (1.6)	10.8 (1.4)	10.6 (0.53)	8.92 (1.1)	8.75 (0.59)	9.03 (0.47)	4.07 (1.1)
biomECG	MIMIC	50,000	5.59 (0.75)	5.22 (0.43)	6.88 (0.51)	8.25 (0.083)	8.09 (1.1)	6.80 (0.72)	6.60 (0.76)	7.10 (0.59)	2.31 (0.26)
biomECG-1k	CODE-15	1,359	14.3 (1.2)	13.1 (1.5)	18.0 (2.4)	19.6 (1.1)	19.6 (2.7)	16.2 (2.4)	15.4 (2.1)	16.4 (1.1)	–

F.2 VISUALIZATION: UMAP OF BIOMETRIC REPRESENTATIONS ON MIMIC

Biometric pretraining produces clearly patient-specific structure in the encoder representation space. Notably, this organization emerges without an explicit contrastive objective: the Siamese training objective alone is sufficient to bring representations from the same patient closer together while separating representations from different patients.

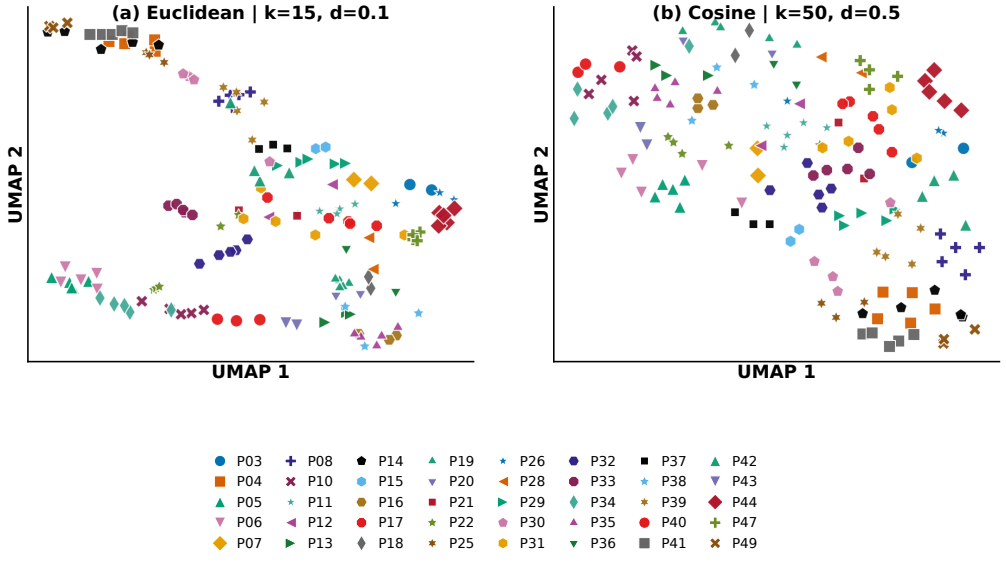

Figure 8: UMAP of 128-dimensional lead-II backbone representations from the biometric encoder pbiomECG-1k, pretrained on 1,359 MIMIC patients. Colors and symbols identify 40 randomly selected patients with repeated recordings; up to their five earliest ECGs are shown. Patients are sampled from the available MIMIC cohort and may include individuals seen during pretraining. UMAP uses Euclidean distance with 15 neighbors and min_dist=0.1 (a), and cosine distance with 50 neighbors and min_dist=0.5 (b). Both projections use identical ECGs and are fitted without patient labels.

G DOWNSTREAM EVALUATIONS

G.1 SEEDS

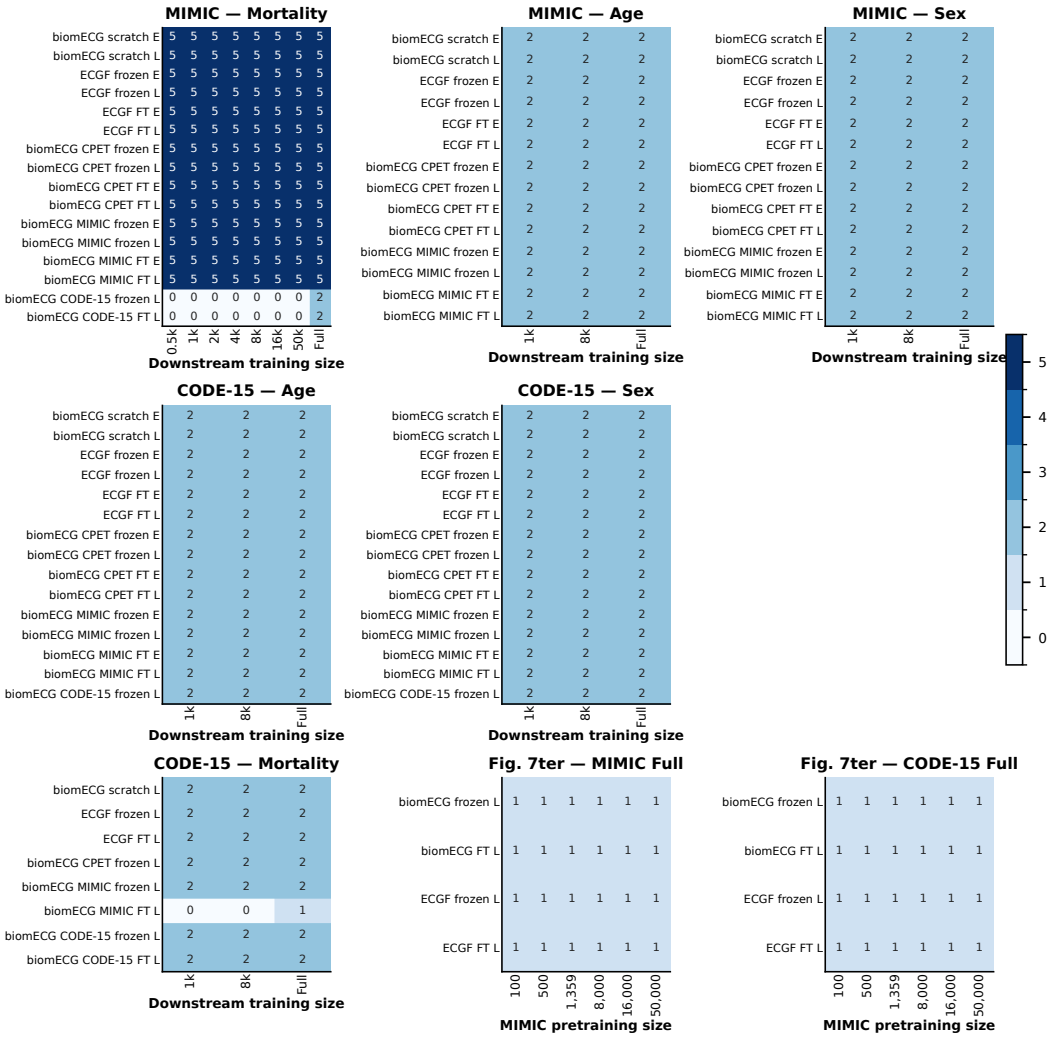

Figure 9: Number of completed seeds per experiment on MIMIC and CODE-15. E: early fusion; L: eight-lead late fusion, counted only when all eight experts are complete. Biometric pretraining uses 1,359 patients. Zero denotes no complete result, including configurations not run.

G.2 AGE PREDICTION BIAIS

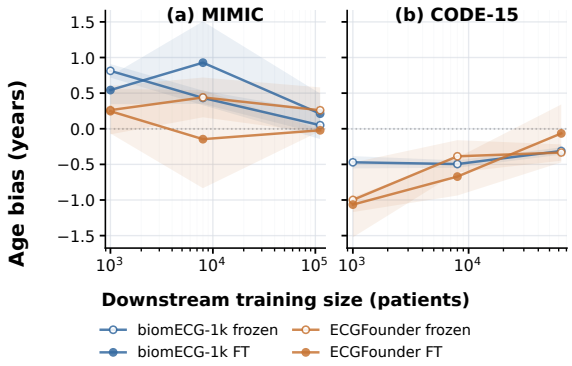

Figure 10: Age prediction bias (predicted minus true age) versus downstream training size, using eight-lead late fusion with frozen or fine-tuned encoders. Positive values indicate overestimation; negative values indicate underestimation. Bias is generally positive on MIMIC and negative on CODE-15. Curves and bands show the mean bias and one SD across seeds. 2 seeds.

G.3 FULL RESULTS

Table 7: MIMIC mortality results (AUROC): mean (SD). Seed counts: Fig. 9. Late fusion uses eight leads; – indicates unavailable results.

Model	Regime	Fusion	500	1,000	2,000	4,000	8,000	16,000	50,000	Full
biomECG scratch	–	Early	0.619 (0.052)	0.642 (0.020)	0.674 (0.013)	0.699 (0.0047)	0.735 (0.0089)	0.767 (0.0056)	0.801 (0.0044)	0.824 (0.0038)
biomECG scratch	–	Late	0.676 (0.020)	0.704 (0.013)	0.732 (0.0063)	0.758 (0.0053)	0.782 (0.0035)	0.803 (0.0076)	0.832 (0.0033)	0.838 (0.0019)
ECGFounder	Frozen	Early	0.761 (0.0073)	0.772 (0.0068)	0.779 (0.0043)	0.790 (0.0059)	0.798 (0.0042)	0.804 (0.0026)	0.815 (0.0017)	0.822 (0.0022)
ECGFounder	Frozen	Late	0.774 (0.0093)	0.785 (0.0080)	0.788 (0.011)	0.798 (0.0069)	0.808 (0.0044)	0.816 (0.0044)	0.834 (0.0040)	0.843 (0.0030)
ECGFounder	FT	Early	0.761 (0.0076)	0.772 (0.0076)	0.779 (0.0035)	0.790 (0.0072)	0.800 (0.0051)	0.808 (0.0026)	0.822 (0.0054)	0.829 (0.0021)
ECGFounder	FT	Late	0.774 (0.0085)	0.786 (0.0071)	0.790 (0.011)	0.797 (0.0055)	0.805 (0.0043)	0.813 (0.0050)	0.824 (0.0042)	0.833 (0.0037)
biomECG-1k CPET	Frozen	Early	0.687 (0.018)	0.717 (0.0072)	0.741 (0.0070)	0.764 (0.0065)	0.785 (0.0065)	0.804 (0.0071)	0.818 (0.0041)	0.825 (0.0027)
biomECG-1k CPET	Frozen	Late	0.755 (0.017)	0.778 (0.0088)	0.786 (0.0079)	0.806 (0.0049)	0.820 (0.0060)	0.829 (0.0033)	0.833 (0.0024)	0.835 (0.0025)
biomECG-1k CPET	FT	Early	0.694 (0.014)	0.711 (0.011)	0.732 (0.0092)	0.757 (0.0036)	0.776 (0.0049)	0.791 (0.0050)	0.810 (0.0037)	0.824 (0.0027)
biomECG-1k CPET	FT	Late	0.747 (0.016)	0.770 (0.0081)	0.778 (0.0073)	0.788 (0.0057)	0.796 (0.0043)	0.806 (0.0051)	0.826 (0.0045)	0.840 (0.0037)
biomECG-1k MIMIC	Frozen	Early	0.728 (0.021)	0.744 (0.014)	0.758 (0.0045)	0.769 (0.0060)	0.785 (0.0046)	0.805 (0.0033)	0.821 (0.0014)	0.828 (0.0015)
biomECG-1k MIMIC	Frozen	Late	0.764 (0.012)	0.780 (0.0073)	0.789 (0.0068)	0.804 (0.0071)	0.820 (0.0051)	0.829 (0.0040)	0.833 (0.0022)	0.835 (0.0025)
biomECG-1k MIMIC	FT	Early	0.720 (0.015)	0.735 (0.013)	0.756 (0.0054)	0.769 (0.0029)	0.780 (0.0045)	0.791 (0.0041)	0.808 (0.0032)	0.824 (0.0036)
biomECG-1k MIMIC	FT	Late	0.759 (0.013)	0.772 (0.0062)	0.778 (0.0037)	0.787 (0.0049)	0.794 (0.0039)	0.803 (0.0038)	0.827 (0.0041)	0.839 (0.0027)
biomECG-1k CODE-15	Frozen	Early	–	–	–	–	–	–	–	–
biomECG-1k CODE-15	Frozen	Late	–	–	–	–	–	–	–	0.831 (0.00028)
biomECG-1k CODE-15	FT	Early	–	–	–	–	–	–	–	–
biomECG-1k CODE-15	FT	Late	–	–	–	–	–	–	–	0.839 (0.0011)

Table 8: MIMIC age results (MAE in years): mean (SD). Seed counts: Fig. 9. Late fusion uses eight leads; – indicates unavailable results.

Model	Regime	Fusion	1,000	8,000	Full
biomECG scratch	–	Early	13.2 (0.20)	11.1 (0.10)	8.31 (0.069)
biomECG scratch	–	Late	13.5 (0.10)	11.1 (0.15)	9.52 (0.087)
ECGFounder	Frozen	Early	10.2 (0.10)	9.41 (0.090)	8.72 (0.012)
ECGFounder	Frozen	Late	10.6 (0.064)	9.79 (0.0052)	9.13 (0.035)
ECGFounder	FT	Early	10.1 (0.16)	9.01 (0.14)	7.77 (0.014)
ECGFounder	FT	Late	10.6 (0.035)	9.57 (0.11)	8.79 (0.11)
biomECG-1k CPET	Frozen	Early	11.7 (0.015)	11.1 (0.098)	10.4 (0.014)
biomECG-1k CPET	Frozen	Late	11.8 (0.0089)	11.1 (0.028)	10.5 (0.059)
biomECG-1k CPET	FT	Early	10.9 (0.078)	9.62 (0.0092)	8.14 (0.033)
biomECG-1k CPET	FT	Late	11.2 (0.017)	10.3 (0.052)	9.42 (0.010)
biomECG-1k MIMIC	Frozen	Early	11.3 (0.10)	10.7 (0.091)	10.2 (0.053)
biomECG-1k MIMIC	Frozen	Late	11.7 (0.026)	11.1 (0.035)	10.6 (0.060)
biomECG-1k MIMIC	FT	Early	10.8 (0.013)	9.89 (0.0076)	8.20 (0.027)
biomECG-1k MIMIC	FT	Late	11.4 (0.011)	10.5 (0.10)	9.49 (0.081)
biomECG-1k CODE-15	Frozen	Early	–	–	–
biomECG-1k CODE-15	Frozen	Late	–	–	–
biomECG-1k CODE-15	FT	Early	–	–	–
biomECG-1k CODE-15	FT	Late	–	–	–

Table 9: MIMIC sex results (AUROC): mean (SD). Seed counts: Fig. 9. Late fusion uses eight leads; – indicates unavailable results.

Model	Regime	Fusion	1,000	8,000	Full
biomECG scratch	–	Early	0.697 (0.0075)	0.824 (0.0049)	0.918 (0.0034)
biomECG scratch	–	Late	0.752 (0.0037)	0.842 (0.0056)	0.901 (0.0076)
ECGFounder	Frozen	Early	0.804 (0.0033)	0.852 (0.0036)	0.883 (0.0036)
ECGFounder	Frozen	Late	0.831 (0.0055)	0.866 (0.0045)	0.891 (0.0046)
ECGFounder	FT	Early	0.820 (0.0018)	0.876 (0.0020)	0.930 (0.0039)
ECGFounder	FT	Late	0.833 (0.0078)	0.873 (0.0071)	0.910 (0.0033)
biomECG-1k CPET	Frozen	Early	0.788 (0.0033)	0.822 (0.0066)	0.845 (0.0047)
biomECG-1k CPET	Frozen	Late	0.824 (0.011)	0.842 (0.0080)	0.858 (0.0039)
biomECG-1k CPET	FT	Early	0.822 (0.0073)	0.875 (0.0030)	0.926 (0.0020)
biomECG-1k CPET	FT	Late	0.829 (0.011)	0.866 (0.0052)	0.903 (0.0037)
biomECG-1k MIMIC	Frozen	Early	0.778 (0.014)	0.804 (0.015)	0.828 (0.011)
biomECG-1k MIMIC	Frozen	Late	0.823 (0.015)	0.841 (0.0045)	0.857 (0.011)
biomECG-1k MIMIC	FT	Early	0.814 (0.012)	0.864 (0.0074)	0.920 (0.0014)
biomECG-1k MIMIC	FT	Late	0.824 (0.013)	0.859 (0.0080)	0.900 (0.0069)
biomECG-1k CODE-15	Frozen	Early	–	–	–
biomECG-1k CODE-15	Frozen	Late	–	–	–
biomECG-1k CODE-15	FT	Early	–	–	–
biomECG-1k CODE-15	FT	Late	–	–	–

Table 10: CODE-15 mortality results (AUROC): mean (SD). Seed counts: Fig. 9. Late fusion uses eight leads; – indicates unavailable results.

Model	Regime	Fusion	1,000	8,000	Full
biomECG scratch	–	Early	–	–	–
biomECG scratch	–	Late	0.668 (0.019)	0.766 (0.015)	0.827 (0.012)
ECGFounder	Frozen	Early	–	–	–
ECGFounder	Frozen	Late	0.754 (0.0015)	0.794 (0.022)	0.829 (0.0094)
ECGFounder	FT	Early	–	–	–
ECGFounder	FT	Late	0.755 (0.015)	0.796 (0.019)	0.824 (0.018)
biomECG-1k CPET	Frozen	Early	–	–	–
biomECG-1k CPET	Frozen	Late	0.705 (0.032)	0.812 (0.019)	0.844 (0.0083)
biomECG-1k CPET	FT	Early	–	–	–
biomECG-1k CPET	FT	Late	–	–	–
biomECG-1k MIMIC	Frozen	Early	–	–	–
biomECG-1k MIMIC	Frozen	Late	0.743 (0.013)	0.816 (0.014)	0.847 (0.0070)
biomECG-1k MIMIC	FT	Early	–	–	–
biomECG-1k MIMIC	FT	Late	–	–	0.813 (–)
biomECG-1k CODE-15	Frozen	Early	–	–	–
biomECG-1k CODE-15	Frozen	Late	0.727 (0.018)	0.807 (0.016)	0.841 (0.0095)
biomECG-1k CODE-15	FT	Early	–	–	–
biomECG-1k CODE-15	FT	Late	0.724 (0.027)	0.772 (0.030)	0.827 (0.017)

Table 11: CODE-15 age results (MAE in years): mean (SD). Seed counts: Fig. 9. Late fusion uses eight leads; – indicates unavailable results.

Model	Regime	Fusion	1,000	8,000	Full
biomECG scratch	–	Early	13.4 (0.61)	11.3 (0.010)	8.68 (0.18)
biomECG scratch	–	Late	13.8 (0.031)	11.5 (0.076)	9.82 (0.0070)
ECGFounder	Frozen	Early	10.2 (0.0041)	9.34 (0.030)	8.66 (0.013)
ECGFounder	Frozen	Late	10.7 (0.11)	9.82 (0.078)	9.20 (0.0016)
ECGFounder	FT	Early	9.93 (0.037)	8.73 (0.092)	7.63 (0.12)
ECGFounder	FT	Late	10.6 (0.13)	9.53 (0.076)	8.60 (0.053)
biomECG-1k CPET	Frozen	Early	11.7 (0.011)	11.1 (0.12)	10.6 (0.096)
biomECG-1k CPET	Frozen	Late	12.5 (0.030)	11.8 (0.10)	11.2 (0.12)
biomECG-1k CPET	FT	Early	11.0 (0.0062)	9.78 (0.014)	8.16 (0.038)
biomECG-1k CPET	FT	Late	11.8 (0.17)	10.7 (0.049)	9.59 (0.057)
biomECG-1k MIMIC	Frozen	Early	11.6 (0.038)	11.0 (0.017)	10.5 (0.033)
biomECG-1k MIMIC	Frozen	Late	12.4 (0.029)	11.7 (0.063)	11.3 (0.15)
biomECG-1k MIMIC	FT	Early	11.0 (0.052)	10.0 (0.060)	8.57 (0.050)
biomECG-1k MIMIC	FT	Late	12.0 (0.0048)	10.9 (0.037)	9.79 (0.12)
biomECG-1k CODE-15	Frozen	Early	–	–	–
biomECG-1k CODE-15	Frozen	Late	13.2 (0.17)	12.4 (0.026)	12.0 (0.032)
biomECG-1k CODE-15	FT	Early	–	–	–
biomECG-1k CODE-15	FT	Late	–	–	–

Table 12: CODE-15 sex results (AUROC): mean (SD). Seed counts: Fig. 9. Late fusion uses eight leads; – indicates unavailable results.

Model	Regime	Fusion	1,000	8,000	Full
biomECG scratch	–	Early	0.678 (0.0011)	0.823 (0.0067)	0.910 (0.00091)
biomECG scratch	–	Late	0.789 (0.0038)	0.862 (0.0038)	0.906 (0.00067)
ECGFounder	Frozen	Early	0.828 (0.0029)	0.869 (0.0022)	0.893 (0.0011)
ECGFounder	Frozen	Late	0.847 (0.0024)	0.883 (0.0048)	0.902 (0.00062)
ECGFounder	FT	Early	0.837 (0.0061)	0.891 (0.0016)	0.931 (0.00062)
ECGFounder	FT	Late	0.847 (0.0015)	0.895 (0.0014)	0.919 (0.0018)
biomECG-1k CPET	Frozen	Early	0.788 (0.0022)	0.823 (0.0018)	0.843 (0.00075)
biomECG-1k CPET	Frozen	Late	0.846 (0.0015)	0.865 (0.0042)	0.873 (0.0010)
biomECG-1k CPET	FT	Early	0.822 (0.00094)	0.881 (0.0064)	0.922 (0.0016)
biomECG-1k CPET	FT	Late	0.854 (0.0050)	0.888 (0.0021)	0.912 (0.0046)
biomECG-1k MIMIC	Frozen	Early	0.783 (0.0048)	0.810 (0.0068)	0.827 (0.0026)
biomECG-1k MIMIC	Frozen	Late	0.843 (0.0011)	0.863 (0.0032)	0.871 (0.0013)
biomECG-1k MIMIC	FT	Early	0.815 (0.0081)	0.867 (0.0018)	0.915 (0.00048)
biomECG-1k MIMIC	FT	Late	0.847 (0.0067)	0.878 (0.00039)	0.906 (0.0036)
biomECG-1k CODE-15	Frozen	Early	–	–	–
biomECG-1k CODE-15	Frozen	Late	0.825 (0.012)	0.841 (0.0095)	0.854 (0.011)
biomECG-1k CODE-15	FT	Early	–	–	–
biomECG-1k CODE-15	FT	Late	–	–	–

Table 13: MIMIC full-cohort mortality AUROC versus MIMIC biometric pretraining size, corresponding to Fig. 3. Eight-lead late fusion; seed 12345 only due to limited computational resources. Entries show AUROC (SD); SD is unavailable for a single seed. ECGFounder references are reused across sizes. Seed counts: Fig. 9.

Model	Regime	MIMIC pretraining patients					
		100	500	1,359	8,000	16,000	50,000
biomECG (MIMIC)	Frozen	0.824 (–)	0.831 (–)	0.835 (–)	0.841 (–)	0.844 (–)	0.847 (–)
biomECG (MIMIC)	FT	0.838 (–)	0.839 (–)	0.840 (–)	0.837 (–)	0.840 (–)	0.843 (–)
ECGFounder	Frozen	0.841 (–)	0.841 (–)	0.841 (–)	0.841 (–)	0.841 (–)	0.841 (–)
ECGFounder	FT	0.831 (–)	0.831 (–)	0.831 (–)	0.831 (–)	0.831 (–)	0.831 (–)

Table 14: CODE-15 full-cohort mortality AUROC versus MIMIC biometric pretraining size, corresponding to Fig. 3. Eight-lead late fusion; seed 12345 only due to limited computational resources. Entries show AUROC (SD); SD is unavailable for a single seed. ECGFounder references are reused across sizes. Seed counts: Fig. 9.

Model	Regime	MIMIC pretraining patients					
		100	500	1,359	8,000	16,000	50,000
biomECG (MIMIC)	Frozen	0.832 (–)	0.835 (–)	0.842 (–)	0.838 (–)	0.845 (–)	0.852 (–)
biomECG (MIMIC)	FT	0.807 (–)	0.816 (–)	0.813 (–)	0.819 (–)	0.818 (–)	0.828 (–)
ECGFounder	Frozen	0.822 (–)	0.822 (–)	0.822 (–)	0.822 (–)	0.822 (–)	0.822 (–)
ECGFounder	FT	0.812 (–)	0.812 (–)	0.812 (–)	0.812 (–)	0.812 (–)	0.812 (–)