MARC: MULTI-SCALE ADAPTIVE RELATION CALIBRATION FOR LONGITUDINAL EHR-BASED HEART FAILURE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting whether heart failure is recorded at a subsequent visit from longitudinal electronic health records (EHRs) requires modeling both cross-visit dependencies and the relations among diagnosis codes. Existing methods exploit temporal information, medical knowledge, and graph structure, yet provide limited control over how these relations are integrated into code representations, graph propagation, and training supervision in a specific prediction context. We propose MARC, an adaptive relation-calibration framework built on the principle of *preserving support while calibrating strength*. MARC comprises three calibration mechanisms: knowledge-aware hierarchical semantic calibration retains code-specific information; state-aware edge calibration controls the propagation strength of observed co-occurrence edges; and class-aware patient relation calibration regularizes within-batch patient relations during training. Evaluated on heart-failure prediction cohorts constructed from MIMIC-III and MIMIC-IV, MARC achieves the highest mean AUROC, F1, and AUPRC and the lowest mean Brier score on both cohorts. A complete factorial study further reveals that the contributions of the three calibration mechanisms depend on the dataset and module combination, rather than combining as fixed additive gains. The source code is available at https://anonymous.4open.science/r/marc-reproducibility-5109/.

1 INTRODUCTION

Heart failure imposes a substantial burden on global health and is associated with recurrent hospitalizations and poor prognosis (Shahim et al., 2023; McDonagh et al., 2021; Heidenreich et al., 2022). Predicting whether heart failure will be recorded at a subsequent visit, given a patient's prior longitudinal clinical records, is therefore an important task in longitudinal EHR modeling. By capturing diagnoses and their timing across repeated visits, EHRs provide a basis for identifying historical patterns that precede a later heart-failure diagnosis.

Artificial intelligence methods have been widely adopted for longitudinal EHR modeling. Temporal models extract cross-visit dependencies from irregularly spaced visit sequences (Choi et al., 2016b; Luo et al., 2020), while knowledge-enhanced and graph–temporal methods further incorporate relations among medical concepts and diagnosis codes (Choi et al., 2017; Lu et al., 2021; Jiang et al., 2024; Poulain & Beheshti, 2024; Li & Li, 2026). Medical hierarchies endow diagnosis codes with shared semantics, whereas co-occurrence graphs estimated from the training cohort capture population-level structural regularities, which have gained superior performance. Both can thus complement longitudinal patient history. However, obtaining or constructing a relation does not by itself determine how it should shape the representation of a specific code, how graph propagation should be conditioned on the patient's current state, or how relations among patient representations should be modeled during training.

The limitation is not that such relations are uninformative, but that their use is not systematically adapted to the prediction context. This problem manifests at three stages of the prediction pipeline. First, coarse higher-level concepts can dilute code-specific information when incorporated too strongly. Second, cohort-level co-occurrence weights are shared across patients and may not be equally predictive under different patient states. Third, patient-wise binary cross-entropy does

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

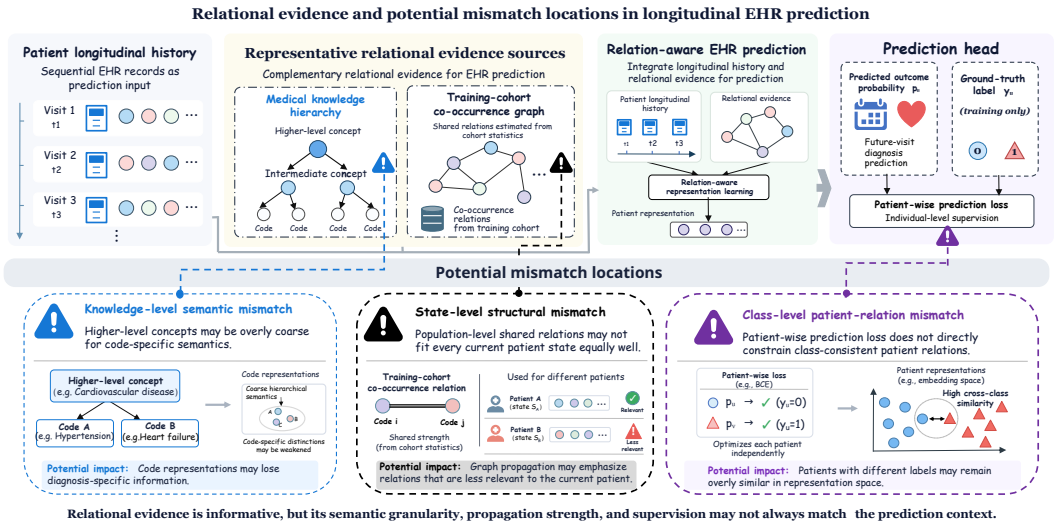

Figure 1: Three relation-use mismatches motivating MARC: coarse hierarchical semantics, patient-invariant cohort-level edge strengths, and patient-wise supervision that does not directly constrain inter-patient relations.

not directly constrain relations among patient representations, allowing patients with different labels to remain highly similar. We refer to these as the *knowledge-level semantic mismatch*, *state-level structural mismatch*, and *class-level patient-relation mismatch*, respectively. Figure 1 illustrates where these mismatches arise. Appendix D further examines their associated model behavior on the validation set; the test set is not used for the design or selection of these analyses.

To address these mismatches, we propose MARC, an adaptive relation-calibration framework following the principle of *preserving support while calibrating strength*. **Knowledge-Aware Hierarchical Semantic Calibration (H)** controls hierarchy-aware updates to diagnosis-code representations; **State-Aware Edge Calibration (E)** conditions the propagation strength of existing co-occurrence edges on the current patient state; and **Class-Aware Patient Relation Calibration (R)** constrains within-batch patient relations using class information and prediction confidence during training only. The three mechanisms target the knowledge-level, state-level, and class-level mismatches, respectively.

Our contributions are threefold:

1. We identify three concrete mismatches in how relations are used in relation-aware heart-failure prediction: insufficient control over the granularity of hierarchical semantics, the lack of patient-state conditioning for shared co-occurrence edges, and the absence of explicit constraints on patient relations under patient-wise supervision. We formulate their calibration jointly as relation-calibration problems.
2. We propose MARC, which applies knowledge-aware hierarchical semantic calibration (H), state-aware edge calibration (E), and class-aware patient relation calibration (R) to code representations, observed co-occurrence edges, and training-stage patient relations, respectively, while preserving the original relation support and leaving the inference path unchanged.
3. We systematically evaluate MARC on MIMIC-III and MIMIC-IV through a complete factorial ablation study and validation-set mechanism analyses, examining the individual contributions, interactions, and design behavior of H, E, and R.

2 Related Work

Temporal and structural modeling of longitudinal EHRs. Longitudinal EHR prediction has evolved from sequence modeling to models that jointly capture temporal and structural dependen-

cies. Doctor AI uses recurrent networks for subsequent-event prediction (Choi et al., 2016a); RETAIN introduces reverse-time attention (Choi et al., 2016b); Dipole models cross-visit relations with attention-based bidirectional recurrent networks (Ma et al., 2017); and HiTANet further incorporates time-aware attention (Luo et al., 2020). Structural approaches extend this line by learning latent EHR relations with Graph Convolutional Transformer (Choi et al., 2020), adapting comorbidity edges to patient history (Qian et al., 2020), or combining graph and temporal modeling, as in GT-BEHRT (Poulain & Beheshti, 2024) and TCEM (Li & Li, 2026). MARC follows this graph–temporal paradigm but focuses on calibrating the strength of existing cohort-level relations according to the current patient state.

Knowledge-enhanced EHR prediction. Medical knowledge has been incorporated through ontologies, knowledge graphs, and language-model semantics. GRAM attentively aggregates a code with its ancestors (Choi et al., 2017), while KAME introduces knowledge-attention for diagnosis prediction (Ma et al., 2018). Later methods combine richer knowledge sources: CGL integrates patient–disease interactions, domain knowledge, and text (Lu et al., 2021); GraphCare constructs patient-specific biomedical knowledge graphs (Jiang et al., 2024); KARE combines multi-source knowledge graphs with retrieval and language-model reasoning (Jiang et al., 2025); and BoxLM unifies ontology- and EHR-derived structures with language-model semantics (Tan et al., 2025). In contrast, MARC does not seek additional knowledge sources, but calibrates how a fixed medical hierarchy updates diagnosis-code representations while preserving code-specific semantics.

3 MARC Framework

In this section, we first formalize the prediction task in Section 3.1 and then present the overall MARC framework in Section 3.2.

3.1 Problem Formulation

We first formalize the heart-failure prediction task. We represent a longitudinal EHR as a sequence of visits, each containing a set of diagnosis codes from vocabulary $\mathcal{C}$. Let patient u have T_u visits, where t indexes visits. We hold out visit T_u as the prediction target and use only the preceding visits as input:

$$\mathcal{X}^{(u)} = \{V_1^{(u)}, \ldots, V_{T_u-1}^{(u)}\}, \tag{1}$$

where $V_t^{(u)} \subseteq \mathcal{C}$ is the set of diagnosis codes recorded at historical visit t. The binary label $y_u \in \{0, 1\}$ indicates whether heart failure is recorded at the held-out visit. This endpoint does not require a first heart-failure diagnosis and does not explicitly model disease progression. Let $\Delta^{(u)}$ denote the historical inter-visit intervals. The model predicts

$$p_u = P\Big(y_u = 1 \mid \mathcal{X}^{(u)}, \Delta^{(u)}\Big). \tag{2}$$

3.2 Framework Overview

MARC formulates relation modeling in longitudinal EHR prediction as an adaptive calibration task: it preserves the original relation support while controlling how hierarchical semantics, cohort-level co-occurrence relations, and patient relations are used under a specific prediction context. The framework operates on a fixed training-cohort diagnosis co-occurrence graph and a fixed medical knowledge hierarchy, and introduces knowledge-aware, state-aware, and class-aware calibration at different stages of the prediction pipeline.

The diagnosis-code graph is constructed only from historical visits of training patients. Each diagnosis code in $\mathcal{C}$ is treated as a node, and diagnosis codes that co-occur within the same historical visit contribute to pairwise co-occurrence counts. The counts are accumulated in both directions without self-loops. After applying the training-set pruning rule, the retained counts are row-normalized to form the weighted adjacency matrix A, yielding the training-cohort diagnosis co-occurrence graph $\mathcal{G} = (\mathcal{C}, \mathcal{E}, A)$, where $\mathcal{E}$ denotes the retained co-occurrence edges. Validation and test visits do not participate in graph construction. The graph support is fixed thereafter; the edge-calibration stage

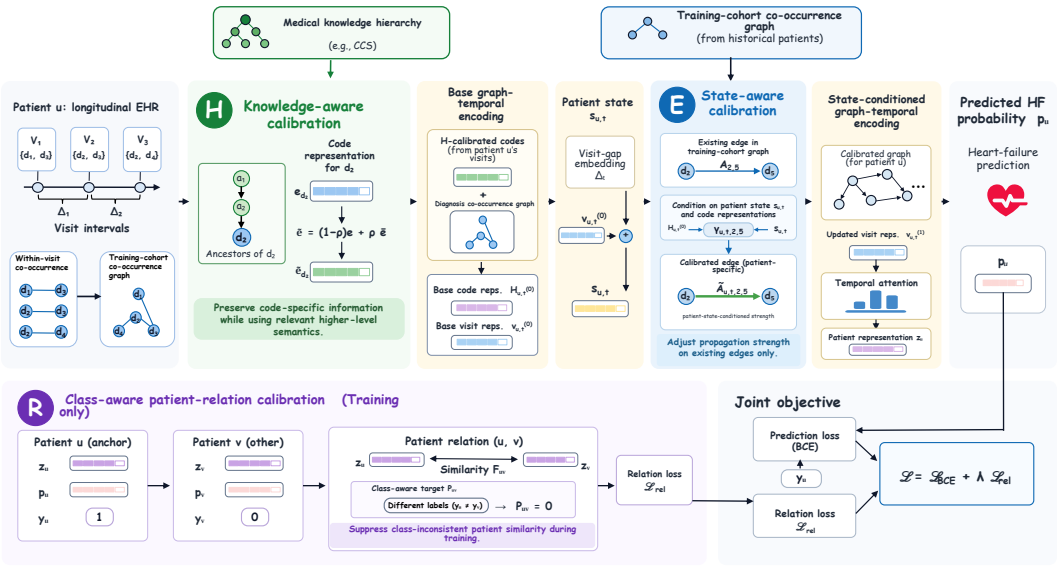

Figure 2: Overview of MARC. H calibrates hierarchical code semantics, E conditions existing co-occurrence edges on patient state, and R regularizes patient relations during training only.

only adjusts the weights of existing edges and does not introduce diagnosis-code edges outside $\mathcal{E}$. Detailed pruning and normalization rules are provided in Appendix B.

Given the fixed training-cohort co-occurrence graph $\mathcal{G}$ described above, MARC adopts a two-stage graph–temporal encoder as its prediction backbone and additionally uses a fixed medical knowledge hierarchy $\mathcal{H}$. **Knowledge-Aware Hierarchical Semantic Calibration (H)** first produces hierarchy-calibrated diagnosis-code representations. The base graph–temporal encoder then combines these representations with A to obtain base code and visit representations, from which the longitudinal state updater constructs the current patient state using the base visit representations and adjacent-visit intervals. **State-Aware Edge Calibration (E)** uses the patient state and base endpoint representations to calibrate the weights of existing co-occurrence edges, yielding a patient–visit-conditioned adjacency matrix. The second-stage graph–temporal encoder then operates on the conditioned adjacency to produce conditional visit representations, which are aggregated by temporal attention into the final patient representation and prediction. **Class-Aware Patient Relation Calibration (R)** provides additional within-batch relational supervision only during training and is removed at inference. Figure 2 illustrates the overall computation flow.

3.2.1 Knowledge-Aware Hierarchical Semantic Calibration

A medical knowledge hierarchy provides shared semantic context that allows diagnosis codes to borrow information from clinically related higher-level concepts. However, such concepts are typically coarser than individual diagnosis codes, and overly strong fusion may weaken code-specific distinctions. H therefore controls both *where* hierarchical information is borrowed from and *how strongly* it is introduced into each diagnosis-code representation.

Let $\mathcal{H}$ denote the fixed medical knowledge hierarchy. For each diagnosis code $i \in \mathcal{C}$, let $e_i \in \mathbb{R}^{d_e}$ denote its learnable base embedding, where d_e is the code-embedding dimension. The candidate concept set of code i is

$$\mathcal{A}_i = \{i\} \cup \mathrm{Anc}_{\mathcal{H}}(i), \tag{3}$$

where $\mathrm{Anc}_{\mathcal{H}}(i)$ contains the ancestor concepts of i along its matched hierarchy path in $\mathcal{H}$. Thus, $\mathcal{A}_i$ contains the diagnosis code itself together with the available higher-level concepts on that path.

For each ancestor concept $a \in \mathrm{Anc}_{\mathcal{H}}(i)$, let $g_a \in \mathbb{R}^{d_e}$ denote its learnable concept embedding. For a candidate $m \in \mathcal{A}_i$, we define its representation for calibrating code i as

$$c_{im} = \begin{cases} e_i, & m = i, \\ g_m, & m \in \mathrm{Anc}_{\mathcal{H}}(i). \end{cases} \tag{4}$$

Both diagnosis-code embeddings and hierarchy-concept embeddings are trainable; their initialization details are provided in Appendix C.

H first determines how strongly each candidate should contribute to the hierarchy-aware representation of code i. For each $m \in \mathcal{A}_i$, the relevance score is

$$\ell_{im} = v_o^\top \tanh(W_o[e_i \| c_{im}] + b_o)\,, \tag{5}$$

where $\|$ denotes concatenation and W_o, b_o, and v_o are learnable parameters. The normalized attention weight is

$$\alpha_{im} = \frac{\exp(\ell_{im})}{\sum_{m' \in \mathcal{A}_i} \exp(\ell_{im'})}, \tag{6}$$

The resulting hierarchy-aware representation is

$$\bar{e}_i = \sum_{m \in \mathcal{A}_i} \alpha_{im} c_{im}. \tag{7}$$

The weights α_{im} therefore determine from which candidates hierarchical information is borrowed.

To control how strongly this information modifies the original diagnosis-code representation, H applies a learnable residual mixture:

$$\widetilde{e}_i = (1-\rho)e_i + \rho\bar{e}_i, \qquad \rho = \sigma(\rho_{\text{logit}}), \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function and ρ_{logit} is a learnable scalar, so $\rho \in (0,1)$. This residual formulation preserves a direct path from the original code representation e_i to the calibrated representation $\widetilde{e}_i$, while ρ controls the strength of the hierarchy-aware update contributed by $\bar{e}_i$.

Collecting the calibrated representations of all diagnosis codes gives

$$\widetilde{\mathbf{E}} = [\widetilde{e}_i]_{i \in \mathcal{C}} \in \mathbb{R}^{|\mathcal{C}| \times d_e}, \tag{9}$$

which denotes the H-calibrated diagnosis-code representation table.

For patient u at historical visit t, let $V_t^{(u)} \subseteq \mathcal{C}$ denote the diagnosis codes recorded at that visit. Given $V_t^{(u)}$, $\widetilde{\mathbf{E}}$, and the fixed training-cohort adjacency matrix A, we apply the first-stage graph–temporal encoder based on the backbone of Li & Li (2026):

$$\left(H_{u,t}^{(0)}, v_{u,t}^{(0)}\right) = \text{TGE}_{\text{base}}\left(V_t^{(u)}, \widetilde{\mathbf{E}}, A\right), \tag{10}$$

where $H_{u,t}^{(0)}$ is the resulting base code-representation matrix and $v_{u,t}^{(0)}$ is the corresponding base visit representation.

3.2.2 State-Aware Edge Calibration

The training-cohort co-occurrence graph provides population-level structural support, but its shared edge weights do not explicitly account for the current patient state. E therefore preserves the observed graph support and conditions only the propagation strength of existing edges on the longitudinal patient state.

Using the base code and visit representations $H_{u,t}^{(0)}$ and $v_{u,t}^{(0)}$ obtained in Section 3.2.1, we first construct the patient state available at visit t. Let $d_{u,t}$ denote the adjacent-visit interval associated with historical visit t. For the first historical visit, we set $d_{u,1} = 0$; for $t > 1$, $d_{u,t}$ denotes the interval between visits $t-1$ and t. Let $\psi_\Delta(\cdot)$ denote the interval encoder. The longitudinal state is updated as

$$s_{u,t} = g_\phi\Big([v_{u,t}^{(0)} \| \psi_\Delta(d_{u,t})], s_{u,t-1}\Big)\,, \qquad s_{u,0} = \mathbf{0}, \tag{11}$$

where $g_\phi(\cdot)$ is the longitudinal state updater parameterized by ϕ. At each visit, the updater combines the current base visit representation and the encoded adjacent-visit interval with the previous patient state. Consequently, $s_{u,t}$ summarizes the longitudinal information available up to visit t and serves as the current patient state for edge calibration.

For an observed directed edge $(i, j) \in \mathcal{E}$, let $h^{(0)}_{u,t,i} = H^{(0)}_{u,t}[i]$ and $h^{(0)}_{u,t,j} = H^{(0)}_{u,t}[j]$ denote its base endpoint representations. E conditions the use of this edge on both endpoint representations and the current patient state:

$$\gamma_{u,t,ij} = \sigma\Big(f_\theta\left([h^{(0)}_{u,t,i} \| h^{(0)}_{u,t,j} \| s_{u,t}]\right)\Big), \tag{12}$$

where $f_\theta(\cdot)$ is the edge-calibration function parameterized by θ, and $\gamma_{u,t,ij} \in (0, 1)$ is the state-aware calibration coefficient.

The corresponding patient–visit-conditioned edge weight is

$$\widetilde{A}_{u,t,ij} = A_{ij}\gamma_{u,t,ij}, \qquad (i, j) \in \mathcal{E}. \tag{13}$$

For edges outside the original support, we define $\widetilde{A}_{u,t,ij} = 0$ whenever $A_{ij} = 0$. Thus, E does not introduce new diagnosis-code relations; it only preserves or attenuates the propagation strength of edges already supported by the training-cohort graph.

Collecting the calibrated edge weights gives the patient–visit-conditioned adjacency matrix $\widetilde{A}_{u,t}$. The second-stage graph–temporal encoder then uses the same H-calibrated code representations together with this conditioned adjacency:

$$v^{(1)}_{u,t} = \mathrm{TGE}_{\mathrm{cond}}\left(V^{(u)}_t, \widetilde{\mathbf{E}}, \widetilde{A}_{u,t}\right), \tag{14}$$

where $\mathrm{TGE}_{\mathrm{cond}}(\cdot)$ denotes the second-stage graph–temporal encoder and $v^{(1)}_{u,t}$ is the resulting conditional visit representation. The sequence of conditional visit representations is subsequently aggregated to form the patient representation used for prediction and class-aware relation calibration.

3.2.3 Class-Aware Patient Relation Calibration

Standard binary cross-entropy supervises patients individually but does not explicitly constrain pairwise relations among patient representations within a mini-batch. Related work has explored relation-consistency regularization between feature relations and class-aware prediction relations (Ding et al., 2025). R therefore introduces a class-aware, confidence-sensitive relation constraint on patient representations during training.

Given the conditional visit representations produced by E, temporal attention aggregates the historical visit sequence of patient u into

$$z_u = \mathrm{Attn}\left(v^{(1)}_{u,1}, \ldots, v^{(1)}_{u,T_u-1}\right), \tag{15}$$

where z_u denotes the final patient representation. The prediction head then produces the heart-failure probability

$$p_u = \mathrm{Cls}(z_u), \qquad p_u \in (0, 1), \tag{16}$$

where $\mathrm{Cls}(\cdot)$ denotes the binary prediction head.

R operates on a mini-batch of B patients during training. Let $k, l \in \{1, \ldots, B\}$ index two patients in the batch. Their representation relation is defined as

$$F_{kl} = \mathrm{ReLU}\left(\cos(z_k, z_l)\right), \tag{17}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. The ReLU operation retains only nonnegative similarity, so F_{kl} measures positive affinity between the learned patient representations.

To construct the class-aware relation target, we define

$$\begin{aligned} q_k &= [1 - p_k, p_k], & S_{kl} &= \mathrm{ReLU}(\cos(q_k, q_l)), \\ M_{kl} &= \mathbf{1}\{y_k = y_l\}, & P_{kl} &= M_{kl}S_{kl}, \end{aligned} \tag{18}$$

where q_k is the binary predictive distribution, S_{kl} the corresponding nonnegative prediction relation, M_{kl} the class mask, and P_{kl} the resulting class-aware target.

For same-class pairs, P_{kl} retains a continuous relation value determined by their predictive distributions, whereas for cross-class pairs the target positive similarity is zero. Thus, R does not force every same-class pair to have similarity one, nor does it require cross-class cosine similarity to become negative.

The relation loss is computed over all unordered patient pairs in the mini-batch:

$$\mathcal{L}_{\text{rel}} = \frac{2}{B(B-1)} \sum_{k<l} \left(F_{kl} - P_{kl}\right)^2 . \tag{19}$$

The condition $k < l$ counts each unordered pair once and excludes self-relations. The prediction-relation branch is not detached; therefore, gradients from $\mathcal{L}_{\text{rel}}$ propagate through both the patient-representation relation F_{kl} and the prediction relation P_{kl}. R is used only during training and therefore does not alter the inference path from historical EHRs to the final prediction probability p_u.

3.2.4 Joint Learning

For a mini-batch of B patients, heart-failure prediction is supervised by binary cross-entropy:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{B} \sum_{k=1}^{B} \left[y_k \log p_k + (1 - y_k) \log(1 - p_k) \right] . \tag{20}$$

The overall training objective combines patient-wise prediction supervision with the class-aware patient relation constraint:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda \mathcal{L}_{\text{rel}}, \tag{21}$$

where $\lambda \geq 0$ controls the contribution of class-aware patient relation calibration to the overall training objective and is specified in Appendix C.1. During training, H and E lie on the prediction path, while R provides additional relational supervision through $\mathcal{L}_{\text{rel}}$. The three calibration modules are therefore jointly optimized with the prediction objective.

4 Experiments

4.1 Performance Comparison on Heart-Failure Prediction

Datasets, Baselines, and Metrics. We evaluate MARC on MIMIC-III (Johnson et al., 2016) and MIMIC-IV v3.1 (Johnson et al., 2023). Both cohorts follow the heart-failure prediction task setting of Li & Li (2026). We compare MARC with three groups of representative methods: temporal models RETAIN (Choi et al., 2016b) and HiTANet (Luo et al., 2020); knowledge- or graph-enhanced methods GRAM (Choi et al., 2017), CGL (Lu et al., 2021), GraphCare (Jiang et al., 2024), and BoxLM (Tan et al., 2025); and TCEM (Li & Li, 2026), which provides the closest graph–temporal backbone to MARC. Detailed baseline configurations and task-specific adaptations are provided in Appendix C.2. We report AUROC, AUPRC, F1, and Brier score.

Experimental Setup. We instantiate $\mathcal{H}$ with the 2015 AHRQ/HCUP multi-level CCS for ICD-9-CM diagnoses (Agency for Healthcare Research and Quality, 2015), covering 91.27% and 90.91% of the diagnosis vocabularies in MIMIC-III and MIMIC-IV, respectively. The patient-state updater is implemented with a GRUCell, and adjacent-visit intervals are clipped to $[0, 15]$ days. MARC is trained with Adam for at most 40 epochs, using validation AUROC for early stopping and checkpoint selection. The effective batch size is 32 for both datasets. Main results and factorial ablations are reported as mean $\pm$ sample standard deviation over five independent runs, and the test set is not used for model selection or mechanism analysis. Further implementation and evaluation details are provided in Appendices A–C.

Results. Table 1 shows that MARC achieves the highest mean AUROC, F1, and AUPRC and the lowest mean Brier score on both cohorts. On MIMIC-III, MARC improves over the strongest baseline GRAM by 0.64 AUROC, 1.68 F1, and 1.88 AUPRC points, while reducing Brier score by 0.33 points. On MIMIC-IV, compared with BoxLM, MARC improves AUROC and AUPRC by 0.73 and 1.61 points and reduces Brier score by 0.27 points, with a smaller 0.20-point F1 gain. Relative to the closest graph–temporal backbone TCEM, MARC also improves all four metrics on both cohorts, supporting the benefit of adaptive relation calibration on top of the shared backbone.

Table 1: Heart-failure prediction on MIMIC-III and MIMIC-IV (mean ± s.d. over five runs).

Dataset	Model	AUROC ↑	F1 ↑	AUPRC ↑	Brier ↓
MIMIC-III	RETAIN	84.22 ± 0.35	69.92 ± 0.95	75.44 ± 1.41	15.44 ± 0.40
	HiTANet	83.71 ± 0.32	70.35 ± 0.91	74.49 ± 1.44	16.07 ± 0.31
	GRAM	85.09 ± 0.21	71.27 ± 1.07	76.76 ± 0.65	14.91 ± 0.20
	CGL	84.32 ± 0.25	69.79 ± 1.59	75.35 ± 0.24	15.35 ± 0.32
	GraphCare	83.06 ± 0.50	67.41 ± 2.00	71.06 ± 1.55	15.99 ± 0.29
	TCEM	83.97 ± 0.86	69.85 ± 1.84	75.91 ± 1.72	15.36 ± 0.43
	BoxLM	84.75 ± 0.21	69.35 ± 2.00	74.43 ± 0.65	15.18 ± 0.21
	MARC	**85.73 ± 0.29**	**72.95 ± 0.53**	**78.64 ± 0.46**	**14.58 ± 0.20**
MIMIC-IV	RETAIN	91.81 ± 0.42	75.09 ± 1.46	78.72 ± 1.61	6.87 ± 0.21
	HiTANet	91.39 ± 0.77	71.44 ± 5.03	78.89 ± 1.96	8.69 ± 2.67
	GRAM	92.57 ± 0.35	74.66 ± 2.02	79.07 ± 1.11	7.01 ± 0.36
	CGL	91.91 ± 0.58	72.50 ± 1.48	77.77 ± 1.70	7.39 ± 0.24
	GraphCare	91.78 ± 0.36	70.90 ± 1.39	76.40 ± 1.23	7.39 ± 0.16
	TCEM	92.43 ± 0.21	76.38 ± 0.52	80.19 ± 1.76	6.70 ± 0.18
	BoxLM	92.84 ± 0.12	78.22 ± 0.18	80.39 ± 0.47	6.56 ± 0.10
	MARC	**93.57 ± 0.45**	**78.42 ± 1.89**	**82.00 ± 1.41**	**6.29 ± 0.33**

4.2 Calibration Mechanism Analysis

Motivation. We examine whether H, E, and R exhibit their intended calibration behavior on the MIMIC-III validation set. The test set is not used for the design or selection of these analyses, and complete controlled results are provided in Appendix D.

Observations. For H, attention-based aggregation performs better overall than uniform aggregation under residual fusion, while direct replacement and residual fusion exhibit metric-dependent trade-offs. The full H model learns a relatively small residual coefficient, $\rho = 0.139 \pm 0.021$, indicating a conservative hierarchy-aware update to the original code representation.

For E, among settings that learn edge gates, state-aware calibration improves all four mean validation metrics over endpoint-only calibration. The learned calibration coefficients remain close to one overall, suggesting that E mainly attenuates existing relations when needed rather than broadly reweighting the graph.

For R, patient-relation calibration primarily reduces positive cross-class similarity, which decreases from 0.480 ± 0.046 without R to 0.336 ± 0.004 with R. This increases the separation between same-class and cross-class relations and reduces relation-target error, indicating that R mainly suppresses class-inconsistent positive similarity rather than simply pulling all same-class patients closer together.

Insights. Together, these results support the intended roles of H, E, and R and show that MARC selectively calibrates relation use rather than requiring large changes to all available relations.

4.3 Ablation Study

Quantitative Analysis. We evaluate all eight combinations of H, E, and R to examine their individual and joint contributions. Tables 2 and 3 report the complete factorial results on the MIMIC-III and MIMIC-IV test sets, respectively. R provides the strongest single-module gains in AUROC and F1 on both datasets. On MIMIC-III, H+R further improves most metrics over either module alone, while the full MARC model achieves the best AUROC, F1, and Brier score. On MIMIC-IV, H+R matches MARC in AUROC, whereas adding E yields the best F1, AUPRC, and Brier score in the full model.

Interpretive Analysis. On MIMIC-III, enabling H, E, or R individually improves several core metrics over the base encoder, with R providing the most pronounced single-module gains. The combinations further show that H and R are complementary, whereas the contribution of E depends

Table 2: Full factorial ablation on MIMIC-III (mean ± s.d. over five runs).

Variant	H	E	R	AUROC ↑	F1 ↑	AUPRC ↑	Brier ↓
Base encoder	–	–	–	83.98 ± 0.86	69.87 ± 1.81	75.91 ± 1.72	15.36 ± 0.43
H only	✓	–	–	84.96 ± 0.48	71.21 ± 1.32	77.61 ± 0.80	14.79 ± 0.30
E only	–	✓	–	84.72 ± 0.77	71.13 ± 0.92	77.61 ± 1.25	14.96 ± 0.41
R only	–	–	✓	85.35 ± 0.43	72.55 ± 0.72	78.66 ± 0.32	14.65 ± 0.21
H + E	✓	✓	–	84.87 ± 0.63	71.38 ± 1.61	77.53 ± 0.83	14.86 ± 0.37
H + R	✓	–	✓	85.59 ± 0.40	72.71 ± 0.85	**78.82 ± 0.64**	14.59 ± 0.35
E + R	–	✓	✓	85.17 ± 0.70	72.51 ± 0.45	78.49 ± 1.07	14.95 ± 0.48
MARC	✓	✓	✓	**85.73 ± 0.29**	**72.95 ± 0.53**	78.64 ± 0.46	**14.58 ± 0.20**

Table 3: Full factorial ablation on MIMIC-IV (mean ± s.d. over five runs).

Variant	H	E	R	AUROC ↑	F1 ↑	AUPRC ↑	Brier ↓
Base encoder	–	–	–	92.66 ± 0.29	76.54 ± 1.12	80.89 ± 1.17	6.56 ± 0.23
H only	✓	–	–	92.69 ± 0.41	76.19 ± 0.57	80.49 ± 0.53	6.68 ± 0.11
E only	–	✓	–	92.17 ± 0.65	76.82 ± 1.51	80.25 ± 1.04	6.72 ± 0.33
R only	–	–	✓	93.23 ± 0.63	77.28 ± 1.16	81.20 ± 1.25	6.58 ± 0.36
H + E	✓	✓	–	91.97 ± 1.76	74.98 ± 2.61	78.08 ± 3.00	7.02 ± 0.72
H + R	✓	–	✓	**93.57 ± 0.46**	77.91 ± 1.79	81.90 ± 1.59	6.37 ± 0.24
E + R	–	✓	✓	93.15 ± 0.96	77.80 ± 0.84	81.96 ± 1.65	6.46 ± 0.22
MARC	✓	✓	✓	**93.57 ± 0.45**	**78.42 ± 1.89**	**82.00 ± 1.41**	**6.29 ± 0.33**

on the modules already enabled. MIMIC-IV exhibits a related but not identical pattern: R again provides the clearest single-module contribution, H has limited effect when used alone but performs better when combined with R, and E mainly improves several metrics when added on top of H+R. These results suggest that the performance of the full MARC model arises from interactions among the three calibration mechanisms rather than from a simple sum of independent module gains.

5 Limitations and Future Work

Our evaluation is limited to MIMIC-III and MIMIC-IV, which originate from the same medical-center system, and the mechanism analyses are primarily based on the MIMIC-III validation set. In addition, the CCS hierarchy does not cover the complete diagnosis vocabulary, and the effects of hierarchy coverage, code frequency, and hierarchy depth remain unexplored. Future work should therefore examine external institutional validation, broader prediction tasks, and stratified analyses of when each calibration mechanism is most beneficial.

6 Conclusion

We proposed MARC for heart-failure prediction from longitudinal EHRs. While preserving relation support, MARC uses knowledge-aware, state-aware, and class-aware mechanisms to calibrate hierarchical semantics, observed co-occurrence edges, and training-stage patient relations, respectively. MARC achieves the strongest overall performance on MIMIC-III and MIMIC-IV among the compared methods. The complete factorial study shows that the contributions of the three modules depend on their combination, while validation-set mechanism analyses further support the effectiveness of controlled calibration of hierarchical semantics, graph propagation, and patient relations. These results suggest that, rather than simply introducing more relations, it is important to control how existing relational evidence is used.

AI Use Statement

Generative AI tools were used to assist with literature discovery and organization, initial drafting and language revision. All AI-assisted text, references, methods, experimental results, and claims

were reviewed and verified by the authors, who take full responsibility for the final content of this work.

Ethics Statement

MIMIC-III and MIMIC-IV are de-identified clinical databases released under controlled data-use agreements. This work is a retrospective binary prediction study using de-identified EHR data and does not provide clinical advice. Both databases originate from the same medical-center system, so the current evaluation does not constitute cross-institutional clinical validation. Prospective validation, subgroup evaluation, and post-deployment monitoring would still be required before clinical use.

Reproducibility Statement

The prediction task and model formulation are defined in Section 3.1, and the main experimental protocol is described in Section 4. Appendix A provides additional details on cohort construction, preprocessing, and evaluation; Appendix B describes the training-cohort diagnosis co-occurrence graph and medical knowledge hierarchy; Appendix C summarizes the implementation and optimization settings of MARC together with baseline task configurations; and Appendix D reports the validation-set mechanism analyses supporting the three calibration modules.

References

Agency for Healthcare Research and Quality. Clinical classifications software (CCS) for ICD-9-CM. Healthcare Cost and Utilization Project (HCUP), 2015. URL https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. 2015 CCS (ICD-9-CM); accessed 2026-09-17.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: predicting clinical events via recurrent neural networks. In Finale Doshi-Velez, Jim Fackler, David C. Kale, Byron C. Wallace, and Jenna Wiens (eds.), *Proceedings of the 1st Machine Learning for Healthcare, MLHC 2016, Los Angeles, CA, USA, August 19-20, 2016*, volume 56 of *JMLR Workshop and Conference Proceedings*, pp. 301–318. JMLR.org, 2016a. URL http://proceedings.mlr.press/v56/Choi16.html.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter F. Stewart. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3504–3512, 2016b. URL https://proceedings.neurips.cc/paper/2016/hash/231141b34c82aa95e48810a9d1b33a79-Abstract.html.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pp. 787–795. ACM, 2017. doi: 10.1145/3097983.3098126. URL https://doi.org/10.1145/3097983.3098126.

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew M. Dai. Learning the graphical structure of electronic health records with graph convolutional transformer. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 606–613. AAAI Press, 2020. doi: 10.1609/AAAI.V34I01.5400. URL https://doi.org/10.1609/aaai.v34i01.5400.

Xi Ding, Lei Wang, Piotr Koniusz, and Yongsheng Gao. Graph your own prompt. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu,

Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/368b5802665d8151a0e2e787bc4ccf94-Abstract-Conference.html.

Paul A Heidenreich, Biykem Bozkurt, David Aguilar, Larry A Allen, Joni J Byun, Monica M Colvin, Anita Deswal, Mark H Drazner, Shannon M Dunlay, Linda R Evers, et al. 2022 aha/acc/hfsa guideline for the management of heart failure: a report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. *Circulation*, 145 (18):e895–e1032, 2022.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=tVTN7Zs0ml.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha A. Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=8fLgt7PQza.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):160035, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240, 2020. doi: 10.1093/BIOINFORMATICS/BTZ682. URL https://doi.org/10.1093/bioinformatics/btz682.

Qing Li and Zehao Li. Two level time-aware network for clinical event prediction. *Eng. Appl. Artif. Intell.*, 164:113259, 2026. doi: 10.1016/J.ENGAPPAI.2025.113259. URL https://doi.org/10.1016/j.engappai.2025.113259.

Chang Lu, Chandan K. Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 3529–3535. ijcai.org, 2021. doi: 10.24963/IJCAI.2021/486. URL https://doi.org/10.24963/ijcai.2021/486.

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 647–656. ACM, 2020. doi: 10.1145/3394486.3403107. URL https://doi.org/10.1145/3394486.3403107.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In

Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017, pp. 1903–1911. ACM, 2017. doi: 10.1145/3097983.3098088. URL https://doi.org/10.1145/3097983.3098088.

Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. KAME: knowledge-based attention model for diagnosis prediction in healthcare. In Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (eds.), *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pp. 743–752. ACM, 2018. doi: 10.1145/3269206.3271701. URL https://doi.org/10.1145/3269206.3271701.

Theresa A McDonagh, Marco Metra, Marianna Adamo, Roy S Gardner, Andreas Baumbach, Michael Böhm, Haran Burri, Javed Butler, Jelena Čelutkienė, Ovidiu Chioncel, John G F Cleland, Andrew J S Coats, Maria G Crespo-Leiro, Dimitrios Farmakis, Martine Gilard, Stephane Heymans, Arno W Hoes, Tiny Jaarsma, Ewa A Jankowska, Mitja Lainscak, Carolyn S P Lam, Alexander R Lyon, John J V McMurray, Alexandre Mebazaa, Richard Mindham, Claudio Muneretto, Massimo Francesco Piepoli, Susanna Price, Giuseppe M C Rosano, Frank Ruschitzka, Anne Kathrine Skibelund, and ESC Scientific Document Group. 2021 esc guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the european society of cardiology (esc) with the special contribution of the heart failure association (hfa) of the esc. *European Heart Journal*, 42(36):3599–3726, 09 2021. ISSN 0195-668X. doi: 10.1093/eurheartj/ehab368. URL https://doi.org/10.1093/eurheartj/ehab368.

Raphael Poulain and Rahmatollah Beheshti. Graph transformers on ehrs: Better representation improves downstream performance. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=pe0Vdv7rsL.

Zhaozhi Qian, Ahmed M. Alaa, Alexis Bellot, Mihaela van der Schaar, and Jem Rashbass. Learning dynamic and personalized comorbidity networks from event data using deep diffusion processes. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3295–3305. PMLR, 2020. URL http://proceedings.mlr.press/v108/qian20a.html.

Bahira Shahim, Chris J Kapelios, Gianluigi Savarese, and Lars H Lund. Global public health burden of heart failure: an updated review. *Cardiac failure review*, 9:e11, 2023.

Yanchao Tan, Hang Lv, Yunfei Zhan, Guofang Ma, Bo Xiong, and Carl Yang. Boxlm: Unifying structures and semantics of medical concepts for diagnosis prediction in healthcare. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/tan25e.html.

APPENDIX

A DATA AND EVALUATION DETAILS

Cohort Construction and Preprocessing. Both datasets use mutually disjoint patient-level training, validation, and test splits. The diagnosis co-occurrence graph is constructed exclusively from historical visits of training patients. Table 4 summarizes the resulting cohort statistics.

For MIMIC-IV, admissions are ordered chronologically and patients with at least two eligible visits are retained. ICD-10-CM diagnosis codes are mapped to ICD-9-CM using the 2018 General Equivalence Mappings (GEM), and unresolved codes are removed. Diagnosis codes are deduplicated

Table 4: Patient-level cohort statistics used in the experiments.

Dataset	Total	Codes	Train	Val.	Test	Val. +	Test +
MIMIC-III	7,493	4,880	6,000	493	1,000	199	367
MIMIC-IV	10,000	5,983	8,000	1,000	1,000	185	185

within each admission and represented as Boolean multi-hot vectors. The final admission is used only to determine the heart-failure label, whereas all preceding admissions form the historical input.

For both cohorts, historical adjacent-visit intervals are clipped to $[0, 15]$ days before encoding. For the first historical visit, the interval is set to zero, consistent with the definition $d_{u,1} = 0$ in Section 3.2.2. The resulting MIMIC-IV cohort contains 10,000 patients and 5,983 diagnosis codes, with an 8,000/1,000/1,000 patient-level train/validation/test split.

For MIMIC-III, we use the patient-level cohort and diagnosis encoding employed in the formal experiments. The resulting cohort contains 7,493 patients and 4,880 diagnosis codes, with a 6,000/493/1,000 train/validation/test split. The same held-out-final-visit prediction protocol is used as for MIMIC-IV.

Evaluation Protocol. All methods use the same patient-level splits. Model selection and early stopping are based exclusively on validation AUROC, and the test set is not used for hyperparameter selection or early stopping. For each independent run, only the checkpoint selected by validation AUROC is evaluated on the test set. The main comparison and complete factorial ablation study report the mean and sample standard deviation over five independent runs.

We report AUROC, AUPRC, F1, and Brier score. AUROC and AUPRC are computed from continuous predictive probabilities. F1 uses a fixed probability threshold of 0.5, without validation-based threshold search. For consistency with the reported tables, all four metrics are presented on a 0–100 scale. Higher values indicate better performance for AUROC, AUPRC, and F1, whereas a lower Brier score indicates better probabilistic prediction quality.

B Relational Structures

B.1 Training-Cohort Diagnosis Co-occurrence Graph

The diagnosis co-occurrence graph is constructed only from historical visits of training patients. For every unordered diagnosis-code pair $\{i, j\}$ appearing in the same visit, both C_{ij} and C_{ji} are incremented. The resulting raw support is therefore bidirectional and contains no self-loops.

For any row whose total raw co-occurrence count exceeds 100, an edge is removed when its preliminary row share satisfies

$$\frac{C_{ij}}{\sum_j C_{ij}} < 0.01. \tag{22}$$

Let C' denote the retained co-occurrence counts after pruning. The adjacency matrix is then obtained through row normalization:

$$A_{ij} = \frac{C'_{ij}}{\sum_j C'_{ij}}. \tag{23}$$

No additional symmetric normalization is applied. Validation and test visits or labels do not participate in graph construction.

Table 5 summarizes the resulting training-cohort graphs.

During MARC training and inference, E computes calibration coefficients only for existing graph support:

$$\widetilde{A}_{u,t,ij} = A_{ij}\gamma_{u,t,ij}. \tag{24}$$

Table 5: Statistics of the training-cohort diagnosis co-occurrence graphs.

Dataset	Nodes	Directed nonzero edges	Self-loops
MIMIC-III	4,880	117,985	0
MIMIC-IV	5,983	143,223	0

If $A_{ij} = 0$, then $\widetilde{A}_{u,t,ij} = 0$. The conditioned adjacency is not renormalized after calibration. Thus, state-aware edge calibration changes the propagation strength of existing relations without introducing new diagnosis-code edges.

B.2 Medical Knowledge Hierarchy

The medical knowledge hierarchy $\mathcal{H}$ is instantiated using the 2015 AHRQ/HCUP multi-level Clinical Classifications Software (CCS) for ICD-9-CM diagnoses (Agency for Healthcare Research and Quality, 2015). It provides a multi-level mapping from diagnosis codes to increasingly coarse clinical concepts.

The hierarchy covers 91.27% of the diagnosis vocabulary in MIMIC-III and 90.91% in MIMIC-IV. For a diagnosis code with a complete matched hierarchy path, the candidate set used by H contains the code itself together with the ancestor concepts on that path. For a code without a complete matched path, the candidate set contains the code itself and a shared root concept.

The hierarchy remains fixed throughout training. H learns how candidate concepts are aggregated and how strongly the resulting hierarchy-aware representation updates the original diagnosis-code representation; it does not modify the hierarchy itself.

C Implementation Details

C.1 MARC Implementation and Optimization

Table 6 summarizes the main configuration of MARC.

Table 6: Core MARC configuration.

Configuration	Value
Code embedding dimension	48
Graph representation dimension	32
Backbone temporal representation dimension	150
State interval-embedding dimension	32
Temporal attention dimension	32
Patient-state dimension	64
Hierarchy-attention dimension	128
Relation-loss weight λ	1
Hierarchy residual initialization	0.1
MIMIC-III physical / effective batch size	32 / 32
MIMIC-IV physical / effective batch size	16 / 32
MIMIC-IV gradient accumulation steps	2

Knowledge-aware hierarchical semantic calibration. Diagnosis-code embeddings have dimension 48. The H attention network receives the concatenation of the code representation and a candidate representation and uses a hidden dimension of 128. Higher-level hierarchy concepts use trainable concept representations shared across the hierarchy-aware aggregation. The global residual coefficient is initialized such that $\rho = 0.1$ and is subsequently optimized jointly with the remaining model parameters.

State-aware edge calibration. The longitudinal state updater is implemented with a GRUCell and uses a 64-dimensional patient state. Adjacent visit intervals are clipped to $[0, 15]$ days, represented by 16 discrete interval bins, and mapped to 32-dimensional interval embeddings.

The edge-calibration function receives two 32-dimensional endpoint representations and the 64-dimensional patient state, yielding a 128-dimensional input. It is implemented as

$$128 \to 32 \to 1,$$

with a ReLU hidden activation followed by a sigmoid output. The same edge-calibration parameters are shared across patients, visits, and observed graph edges.

Class-aware patient relation calibration. R uses cosine similarity followed by ReLU to construct nonnegative patient-representation relations. The relation loss is computed over unordered patient pairs and excludes diagonal self-relations. Prediction probabilities are represented as $[1-p, p]$ before constructing the prediction relation. The prediction-relation branch remains differentiable in the full model, so gradients from $\mathcal{L}_{\text{rel}}$ propagate through both representation and prediction relations. R is removed entirely during validation and testing.

Optimization and computational environment. MARC is optimized using Adam (Kingma & Ba, 2015) with zero weight decay and no dropout for at most 40 epochs. The learning rate is set to 0.01 at epoch 1, 0.001 at epoch 2, 0.0001 for epochs 3–19, and 0.00001 from epoch 20 onward.

Early stopping is based on validation AUROC, with patience 5 and a minimum improvement of 10^{-4}. For each independent run, the checkpoint with the highest validation AUROC is selected and evaluated once on the test set.

Experiments were conducted using Python 3.8 and PyTorch 1.8.1 with CUDA 11.1 on an NVIDIA GeForce RTX 3090 GPU. The complete MARC model contains 806,175 trainable parameters.

C.2 Baseline Configuration and Task Adaptation

All comparison methods use the same patient-level train/validation/test splits and validation-AUROC-based checkpoint selection. Because the original methods differ in task formulation and available modalities, we use task-specific configurations that preserve their central modeling components while restricting patient-level input to information available in our heart-failure prediction setting. Table 7 summarizes these configurations.

Table 7: Baseline inputs and task-specific configurations.

Method	Inputs used in our evaluation	Task-specific configuration
RETAIN	Historical diagnosis codes	Reverse-time recurrent attention with a binary prediction output.
HiTANet	Historical diagnosis codes and adjacent visit intervals	Time-aware visit modeling with a binary prediction output.
GRAM	Historical diagnosis codes and the fixed CCS hierarchy	Hierarchy-aware diagnosis representation followed by binary heart-failure prediction.
CGL	Historical diagnosis codes and the training-cohort co-occurrence graph	Diagnosis-only configuration retaining code-graph propagation, temporal modeling, and visit attention; auxiliary patient text is not used.
GraphCare	Historical diagnosis codes and the training-cohort co-occurrence graph	Diagnosis-only configuration; additional text, laboratory, prescription, and other optional patient modalities are not used.
TCEM	Historical diagnosis codes, visit timing, and the training-cohort co-occurrence graph	The two-stage graph–temporal backbone is retained while MARC-specific calibration modules are disabled.
BoxLM	Historical diagnosis codes, adjacent visit intervals, the CCS hierarchy, and medical-concept semantics	Adapted from next-visit multilabel diagnosis prediction to binary heart-failure prediction while retaining its box representation, structure–semantic fusion, temporal patient representation update, and volume-based scoring.

For BoxLM, medical-concept semantic representations are generated from a frozen BioBERT model (Lee et al., 2020) and are used only for diagnosis-code and hierarchy-concept semantics. No patient

clinical text, patient labels, laboratory tests, prescriptions, or other patient-level modalities are introduced through this semantic representation.

These task-specific configurations ensure that all methods operate on information available before the held-out final visit while preserving the main modeling mechanism of each baseline.

D Additional Mechanism Analysis

This appendix provides additional validation-set evidence for the calibration behavior discussed in Section 4. All analyses are conducted on the MIMIC-III validation set and report the mean and sample standard deviation over five independent runs. The test set is not used to design or select these analyses. The complete factorial ablation results are reported in the main paper; here, we vary design factors directly associated with H, E, and R while holding the remaining configuration fixed.

D.1 Knowledge-Level Analysis

We compare four ways of incorporating hierarchical information: no hierarchy, uniform aggregation followed by residual fusion, attention aggregation with direct replacement, and the full attention-plus-residual design.

Table 8: Hierarchy-calibration analysis on the MIMIC-III validation set. Values are mean ± sample standard deviation over five independent runs.

Setting	AUROC ↑	AUPRC ↑	F1 ↑	Brier ↓
No hierarchy	85.77 ± 0.54	82.40 ± 0.70	73.88 ± 0.24	14.58 ± 0.16
Uniform aggregation + residual	85.81 ± 0.83	82.50 ± 0.48	73.83 ± 1.07	14.64 ± 0.38
Attention aggregation + direct replacement	**86.33 ± 0.39**	82.47 ± 0.78	74.18 ± 0.42	**14.37 ± 0.18**
Attention aggregation + residual (full H)	86.23 ± 0.21	**82.63 ± 0.39**	**74.75 ± 1.15**	14.44 ± 0.27

Under residual fusion, attention-based aggregation performs better overall than uniform aggregation. Under the same attention aggregation, direct replacement and residual fusion exhibit metric-dependent trade-offs: direct replacement gives a higher mean AUROC and lower Brier score, whereas residual fusion gives higher mean AUPRC and F1. These results support separating the choice of hierarchical information source from the strength of the hierarchy-aware update.

The learned global residual coefficient of full H is

$$\rho = 0.139 \pm 0.021.$$

Because the hierarchy-aware aggregate $\bar{e}_i$ contains both the diagnosis code itself and higher-level concepts, ρ should not be interpreted as the final proportion assigned specifically to ancestor concepts. Instead, it indicates that the selected models use a relatively conservative hierarchy-aware update.

D.2 State-Level Analysis

We compare no edge calibration, endpoint-only edge calibration, and the full state-aware edge calibration. The latter two both learn edge gates; their difference is whether the explicit patient state $s_{u,t}$ is included.

Table 9: Edge-calibration analysis on the MIMIC-III validation set. Values are mean ± sample standard deviation over five independent runs.

Setting	AUROC ↑	AUPRC ↑	F1 ↑	Brier ↓
No edge calibration	**86.51 ± 0.29**	**83.00 ± 0.54**	74.38 ± 0.85	**14.34 ± 0.31**
Endpoint-only edge calibration	85.87 ± 0.40	82.56 ± 0.40	74.39 ± 0.93	14.55 ± 0.13
State-aware edge calibration (full E)	86.23 ± 0.21	82.63 ± 0.39	**74.75 ± 1.15**	14.44 ± 0.27

Among the two settings that learn edge gates, adding explicit patient state improves all four mean validation metrics relative to endpoint-only calibration, indicating that the patient state provides conditioning information beyond the endpoint representations. At the same time, the no-edge-calibration setting remains stronger on mean AUROC, AUPRC, and Brier score, while full E gives the higher mean F1. This result is consistent with the interaction-dependent contribution of E observed in the main factorial study.

We additionally examine the learned calibration coefficients on observed validation edges. The γ values of the full state-aware model remain close to one overall, indicating that many training-cohort relations are largely preserved, whereas removing explicit patient state leads to stronger attenuation. Since

$$\widetilde{A}_{u,t,ij} = A_{ij}\gamma_{u,t,ij}, \tag{25}$$

E changes only the strength of existing edges and introduces no relations outside the original graph support.

D.3 Class-Level Analysis

We first compare patient-relation structure with and without R while holding the remaining configuration fixed.

Table 10: Patient-relation structure on the MIMIC-III validation set. Values are mean ± sample standard deviation over five independent runs.

Setting	Same-class relation	Cross-class relation	Separation gap	Relation MSE	Top-5 same-class neighbors
Without R	0.662 ± 0.051	0.480 ± 0.046	0.182 ± 0.010	0.209 ± 0.016	0.691 ± 0.015
With R	0.581 ± 0.009	0.336 ± 0.004	$\mathbf{0.245 \pm 0.006}$	$\mathbf{0.148 \pm 0.003}$	$\mathbf{0.711 \pm 0.010}$

With R, the mean positive similarity among same-class patients does not increase. The larger change is the reduction in positive cross-class similarity, which increases the separation between same-class and cross-class relations, reduces relation-target error, and increases the proportion of same-class nearest neighbors. R is therefore better interpreted as suppressing class-inconsistent positive similarity rather than simply pulling all same-class patients closer together.

We further examine three design choices in the class-aware relation target: gradient propagation through the prediction relation, confidence-sensitive same-class relations, and the class-aware mask.

Table 11: Controlled analysis of the class-aware patient relation target on the MIMIC-III validation set. Values are mean ± sample standard deviation over five independent runs.

Relation target	Validation AUROC ↑	Validation F1 ↑
Full objective	$\mathbf{85.77 \pm 0.54}$	$\mathbf{73.88 \pm 0.24}$
Stop-gradient prediction relation	85.53 ± 0.53	72.74 ± 0.58
Label-only relation	85.38 ± 0.54	73.02 ± 1.58
Without class-aware mask	85.00 ± 0.75	72.44 ± 1.00

The full relation objective achieves the highest mean validation AUROC and F1 among these controlled variants. Removing gradient propagation through the prediction relation, replacing the continuous prediction relation with a label-only target, or removing the class-aware mask all weakens the observed validation performance. Because these analyses are restricted to the MIMIC-III validation set, we interpret them as mechanism evidence rather than as cross-dataset conclusions.