European Journal of Prosthodontics and Restorative Dentistry (2026) 34 (7s), 664–679

CIPHER-Twin: An Invariant and Conformal Digital-Twin Graph for Robust Risk and Time-to-Event Prediction on Heterogeneous EHR Cohorts

Keywords
robust machine learning, dataset shift, digital twin, electronic health records, conformal prediction, anchor regression, federated learning, survival analysis, calibration, social determinants of health

Authors
[1*]Suman
Department of Computer Science & Engineering, University Institute of Engineering and Technology, Maharshi Dayanand University, , Rohtak, 124001, Haryana, India
Emaii: punia.suman@gmail.com,

[2]Yudhvir Singh
Department of Computer Science & Engineering, University Institute of Engineering and Technology, Maharshi Dayanand University, , Rohtak, 124001, Haryana, India
Email: yudhvirsingh@rediffmail.com

[3]Neha Gulati
University Business School, Panjab University, , Chandigarh, India
Email: nehagulati_pu@rediffmail.com

Abstract
Electronic health record (EHR) models that move between hospitals face strong dataset shift, weak or delayed labels, and tight requirements on calibration, fairness, and privacy. Study presents CIPHER–Twin, a digital-twin inspired but engineering-focused framework that treats these issues as a robustness and reliability problem on heterogeneous tabular EHR data. Each case is encoded as a typed patient–phenotype graph with nodes for patient, social determinants of health (SDoH), organs, biomarkers, and therapies, and then passed through a lightweight encoder that supports self-supervised pretraining, federated simulation, and downstream multi-task prediction. The encoder uses graph-masked self-supervision to learn biomarker embeddings conditioned on SDoH, and then applies an anchor-based invariance objective to reduce environment-specific shortcuts under cohort shift. Per-label conformal thresholds convert probabilistic outputs into set-valued risk predictions with finite-sample coverage guarantees, while a discrete-time survival head supports time-to-event modelling. A simple FedAvg loop with clipped and noise-perturbed client updates simulates privacy-aware federated training without moving raw data, and an extremely compact student model distilled from the full encoder enables low-latency deployment. Experiments on four public cohorts (Framingham cardiovascular, UCI Diabetes, UCI CKD, and Heart Failure Clinical Records) show strong discrimination on the test split (AUROC: CVD 0.999, T2DM 0.9998, CKD 0.973; AUPR:
CVD 0.922, T2DM ≈ 1.000, CKD 0.331) while maintaining near-nominal positive-class coverage with compact set-valued outputs (coverage ≈ 0.88–0.92; average set size ≈ 0.95). Group-wise calibration by sex shows small AUC gaps for CVD and T2DM (≤ 0.003). A tiny student distilled from the full encoder reproduces teacher embeddings with cosine similarities peaked near 1.0 while reducing parameters from ~ 9,800 to ~ 1,150 (~ 8.5×), which supports embedded or edge deployment. Survival modelling on the heart-failure cohort reaches a modest C-index (≈ 0.42), consistent with the cohort size and limited longitudinal depth, and serves as a stress test for robustness under sparse time-to-event information. Overall, CIPHER–Twin acts as an end-to-end engineering recipe for robust EHR modelling under dataset shift: it aligns typed tabular representations, invariant objectives, and conformal prediction with a federated-ready training loop and fully auditable exports for reproducible evaluation.

Received:20-06-2026
Revised:25-07-2026
Accepted:03-08-2026
Doi:10.1922/ejprd.v34i7s.1685

1. INTRODUCTION

Digital twins in healthcare AI aim to provide patient-specific forecasting and control by combining data-driven models with physiological knowledge. Yet routinely collected electronic health records (EHRs) remain sparse, ir-regular, and heterogeneous: labels are often weak or delayed, feature distri-butions vary across institutions, and deployment must respect strong con-straints on uncertainty, fairness, and privacy [1, 2]. Classical supervised models perform well when training and deployment settings match, but both discrimination and calibration degrade under dataset shift, including covari-ate and label shift [3, 4]. In parallel, many digital twin candidates for chronic disease management treat physiology only implicitly, so they do not fully use knowledge about glucose–insulin dynamics, blood-pressure regulation, or kid-ney function [5, 6, 7]. These gaps limit the reliability of risk prediction and survival modeling for cardiovascular disease, type 2 diabetes, and chronic kidney disease [8, 9, 10, 11, 12, 13, 14].

CIPHER-Twin (Conformal, Invariant, PHysics-guided, and Efficient Representation) addresses these challenges by building a digital twin representation on top of a typed, multi-entity graph that links patients, organs, biomarkers, ther-apies, and social determinants of health. A lightweight encoder maps this graph into a shared embedding space through self-supervised pretraining fol-lowed by task-specific adaptation [15, 16, 17, 18, 19, 20, 21]. The framework focuses on two common and clinically relevant targets: (i) multi-label risk prediction for co-occurring chronic conditions and (ii) discrete-time survival for time-to-event outcomes such as major adverse cardiovascular events and chronic kidney disease which supports stable optimization, does not require a proportional hazards assumption, and remains compatible with concordance-based evaluation and conformal calibration.

There are also challenges to widespread adoption such as privacy and data governance. Typically, EHRs are contained within a single institution and are not easily transferable between locations. CIPHER-Twin therefore adopts a federated-ready training strategy that simulates multi-site learning through per-site optimization and FedAvg-style aggregation [37, 38, 39, 40, 41, 42]. Gradient clipping and additive noise are in the spirit of differentially private learning to reduce the leakage of information about the individual records [43, 44, 45, 46, 47, 48, 49]. The framework also features a distilled version of the encoder, which takes only social determinants of health as input, to minimize the use of constrained hardware while maintaining the majority of the accuracy of the full graph-based encoder and affording low latency triage.

Within this setting, the study pursues four specific and testable objectives that align with the reported results:

- Design a typed patient graph and encoder for tabular EHR that sup-port both multi-label chronic disease risk prediction and discrete-time survival, and compare their performance with standard machine learn-ing and survival baselines [8, 9, 10, 11, 12, 13, 14].
- Evaluate an anchor-based invariance objective that uses site-level or cohort-level information as anchors, and quantify its effect on out-of-site generalization under dataset shift [3, 24, 25, 26, 27].
- Apply conformal prediction on top of the digital twin outputs for both classification and survival tasks, and measure coverage, set size, and calibration quality under realistic split conditions [28, 29, 30, 31, 32,
- 33, 34, 35, 36, 27].
- Assess a federated-ready training scheme with gradient clipping and noise, and a distilled encoder based only on social determinants of health, in terms of predictive performance and potential for privacy-preserving, resource-aware deployment [43, 37, 44, 46, 47, 39, 40, 48, 49].

Together, these components aim to move digital twin models for chronic disease management closer to robust, interpretable, and privacy-aware use in real-world clinical settings.

2. RELATED WORK

2.1.Risk and time-to-event prediction in chronic disease

Modern risk stratification is a development of time-to-event models which convert observations made on a longitudinal basis into actionable prognos-tics. While still widely used for their ease of use and interpretability, classical Cox models might be prone to failure in the presence of non-proportional hazards, high dimensionality, or nonlinearity [13, 9]. Machine-learning and deep-learning methods such as random survival forests, boosting, and neural hazards, have in turn shown

deterioration [8, 22, 23, 12, 13, 14]. Physics-inspired penalties and an anchor-style invariance objective bias representation learn-ing toward stable clinical mechanisms and reduce dependence on site-specific shortcuts that can arise from differences in coding, feature availability, or follow-up patterns [24, 25, 26, 27].

Uncertainty quantification plays a central role in the digital twin design. Modern neural predictors often produce overconfident probabilities, which complicates risk communication and shared decision-making [4]. To address this issue, CIPHER-Twin uses conformal prediction to transform probabilis-tic scores into set-valued outputs that achieve target coverage on held-out data [28, 29, 30, 31, 32, 33]. In this setting, each patient receives a set of possible labels or a prediction band for event risk, together with a cover-age level that does not rely on strong distributional assumptions. Recent advances in conformal methods for causal and robust prediction motivate this choice in the presence of distribution shift and time-to-event outcomes [26, 34, 35, 36, 27].

Time-to-event modeling uses a discrete-hazard parameterization that fits naturally with EHR follow-up data and aligns with common survival metrics [8, 9, 10, 12, 14]. Let K denote the number of discrete time bins and let hk(x) ∈ (0, 1) be the predicted hazard for a patient with covariates x at bin

k. The implied survival function takes the form

$$S_k(x) = \prod_{j=1}^{k} (1 - h_j(x)), \qquad k = 1, \ldots, K, \tag{1}$$

improvements in discrimination for cardiovas-cular and diabetes outcomes, and have the ability to deal with complex cen-soring patterns [8]. For the task of kidney disease, sequence models (such as Transformers) offer better generalization for variable time intervals in EHRs [14]. However, systematic evidence highlights that institutional and care pathway calibration, transportability, and transparency are still challenges [12]. These gaps drive the architectures that combine powerful discrimination with explicit mechanisms for robustness and uncertainty.

2.2.Representation learning for EHR/IoMT (tabular, temporal, graph)

Representation learning has moved from providing embeddings at a visit-level to sequence models, to relational encoders that take advantage of clin-ical structure. Early deep EHR research found the ability of unsupervised representations to predict a variety of outcomes from a wide range of in-puts [2] and surveys of temporal modeling capabilities document a quick proliferation of abilities [1]. More recently, graph-based approaches model the relationships between patients, diagnoses, biomarkers and treatments; they enhance the prediction and retrieval of cohorts of patients under irreg-ular sampling and missingness [18, 19]. At the cyber-physical edge, IoMT streams add physiological and behavioral signals, which are then used to per-form real-time anomaly detection and triage in a distributed setting using hybrid CNN/LSTM or GNN encoders [50, 38]. While the shift to typed, multimodal graphs is a way towards the digital-twin goals, harmonization, privacy and compute-efficiency are still important practical aspects to con-sider.

2.3.Physiology-guided ML and neural (C)DEs

Physiology-guided learning enforces mechanistic priors based on haemo-dynamics, metabolism and organ function on data-driven models. Examples of such surrogates are ML models of coronary flow physiology, which are clin-ically faithful yet less invasive using imaging information, and physics-based flow models, which can be used to inform cardiovascular decision support at scale. Meanwhile, neural differential equations (ODEs/CDEs) provide continuous-time latent dynamics, which are well adapted to irregular sam-pling and can be used as a means to connect physiology with flexible func-tion approximation in real-world monitoring (e.g., ECG-based guidance) [54]. All of these strands converge towards digital-twin pipelines that incorporate mechanistic plausibility without compromising the statistical efficiency.

2.4.Causal invariance and conformal prediction

Under institutional change, generalization must be achieved, which means that learning mechanisms have to be stable across environments. Invari-ant risk minimization and anchor regression are formalizations of this goal, which reduce or remove environment-specific variation, and are well-suited for multi-site EHR given the variability in coding practices and feature avail-ability across sites [24, 25]. Conformal prediction provides distribution-free, finite-sample coverage in classification and risk stratification, and has been introduced and methodologically developed in safety-critical areas where it is beneficial to abstain from a prediction in a calibrated manner [28, 29, 30]. The combination – learning invariant structure and wrapping predictions with coverage guarantees – is directly geared towards trust under dataset shift.

2.5.Federated learning and differential privacy

The clinical data are often not centralizable and federated learning is a viable path towards collaborative modeling. Performance without raw data sharing is achieved with communication-efficient aggregation across sites [37] and principled leakage control is related to gradient clipping and noise injec-tion, which are part of the well-studied toolbox of differential privacy [43]. Federated setup also exacerbates the heterogeneity in healthcare networks and IoMT ecosystems, leading to personalized variants and graph-aware ap-proaches that are more relevant to local cohorts [19]. These ingredients – federation, privacy and personalisation – are key to the process of making digital-twin prototypes deployable, auditable systems.

3. METHODOLOGY

3.1.Data, Harmonization, and Problem Setup

This study uses four public, de-identified EHR cohorts to test CIPHER–Twin under heterogeneous feature spaces, missingness patterns, and outcome definitions. The Framingham cardiovascular cohort contains 4,240 records and 15 variables, the Diabetes 130-US hospitals cohort contains 101,766 en-counters and 50 variables, the UCI CKD cohort contains 400 records and 25 variables, and the Heart Failure Clinical Records cohort contains 299 sub-jects and 13 variables. These cohorts support two tasks: multi-label chronic disease risk prediction for cardiovascular disease (CVD), type-2 diabetes mel-litus (T2DM), and chronic kidney disease (CKD), and discrete-time survival modelling on the heart-failure cohort. The pipeline uses a fixed row-level split of 60% training, 20% calibration, and 20% test data. The calibration split is reserved for conformal thresholding, scheduler monitoring, and early stopping, while the test split remains untouched until final evaluation. The diabetes cohort includes repeated admissions; therefore, the study treats the experiment as an end-to-end robustness demonstration rather than a patient-deduplicated clinical benchmark.

Table 1: Compact summary of cohorts, targets, and harmonized concepts used by CIPHER–Twin

.

Cohort	Size	Main role	Representative harmonized concepts
Framingham	4,240, 15 vars	CVD proxy-risk source	age, sex, smoking, blood pressure, cholesterol, glucose
Diabetes 130-US	101,766, 50 vars	T2DM proxy-risk source	demographics, admissions, diagnoses, therapies, glucose-related fields
CKD	400, 25 vars	CKD proxy-risk source	renal biomarkers, urine findings, blood pressure, anemia indicators
Heart Failure	299, 13 vars	Survival source	follow-up time, death event, ejection fraction, serum creatinine, serum sodium

A deterministic harmonization pipeline first maps all columns to a com-mon snake_case vocabulary, converts age buckets to numeric midpoints, maps sex and smoking fields to binary or missing-aware numeric codes, and attaches an environment tag E ∈ {framingham, diabetes, ckd, heartfail}. Administrative identifiers, survival-only fields, and training-only variables are excluded from the biomarker encoder and retained only for splitting, supervision, and evaluation. Numeric biomarkers keep missing values dur-ing preprocessing to avoid early cross-sample imputation. Robust scaling is applied only inside the training pipeline as

$$x^\star = \frac{x - \mathrm{median}(x)}{\mathrm{IQR}(x)}, \qquad \mathrm{IQR} = Q_{0.75} - Q_{0.25}, \tag{2}$$

with a small clamp on IQR to avoid division by zero. The harmonized feature space is mapped to a compact ontology with five node types: patient, sdoh, organ, biomarker, and therapy. This ontology ensures that the representation is auditable as each feature consumed is related to a clinical concept.

The multi-label target $Y(x) = (Y_{\mathrm{CVD}}, Y_{\mathrm{T2DM}}, Y_{\mathrm{CKD}}) \in \{0,1\}^3$ follows a transparent weak-label convention:

$$Y_{\mathrm{CVD}} = \mathbb{I}\{E = \texttt{framingham}\}, \quad Y_{\mathrm{T2DM}} = \mathbb{I}\{E = \texttt{diabetes}\}, \quad Y_{\mathrm{CKD}} = \mathbb{I}\{E = \texttt{ckd}\}. \tag{3}$$

This convention produces output space that is shared, without making out-of-scope diagnoses. It is noisy-by-design, with the framework mitigating cohort over-fitting via graph-masked pretraining, anchor-invariance and split conformal calibration. To assure continuity of follow-up, the heart-failure cohort provides follow-up time Ti and event indicator δi. For K time bins and hazard hk(x) ∈ (0, 1), the predicted survival curve is

$$S_k(x) = \prod_{j=1}^{k} (1 - h_j(x)), \qquad k = 1, \ldots, K. \tag{4}$$

3.2.Typed Graph Encoder

For every subject or encounter i, CIPHER–Twin constructs a lightweight typed graph

$$\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \phi, \psi),$$

The types of nodes are {patient, sdoh, organ, biomarker, therapy} and the types of edges are {has, therapy_affects, exposure_affects}. A patient node is connected to the organ nodes (heart, kidney, pancreas) and to the nodes with observed biomarkers and nodes with available therapies. SDoH nodes contain age, sex.

However, it is important to keep the smoking status separate from the biomarkers, as this will make sub-group auditing and student distillation simple. Therapy-to-biomarker and SDoH-to-organ links are interpretable priors not dense message passing constraints. Let b ∈ RM denote the biomarker vector and m ∈ {0, 1}M denote the observation mask, where mj = 1 means biomarker j is observed. The encoder maps each observed biomarker through a shared MLP f (·), averages only
observed embeddings, and fuses this summary with an SDoH embedding:

$$z_b = \frac{1}{\max(1, \sum_{j=1}^{M} m_j)} \sum_{j=1}^{M} m_j f(b_j), \qquad z_s = g(s), \qquad z = h([z_b \| z_s]). \tag{5}$$

Here, g(·) and h(·) are shallow MLPs, and [· ·] denotes concatenation. This design keeps computation linear in the number of observed biomarkers, avoids imputation leakage, and allows missingness to act as useful signal. The pro-totype therefore uses the typed graph as both a semantic alignment layer and an execution graph; denser message passing can replace the masked ag-gregation when richer longitudinal or relational data become available.

Optional physiological penalties encode simple clinical priors for glucose–insulin behavior, kidney function, and blood-pressure regulation. These penalties are kept diagnostic or weakly weighted in the main experiments because the available cohorts are mostly cross-sectional. For edge deploy-ment, a compact SDoH-only student qω(s) learns to reproduce the teacher embedding z by minimizing

$$\mathcal{L}_{\mathrm{distill}} = \left\| \frac{q_\omega(s)}{\|q_\omega(s)\|_2} - \frac{z}{\|z\|_2} \right\|_2^2.$$

The student uses only age, sex, and smoking status and supports low-latency triage when biomarker availability is limited.

3.3.Training Objectives

3.3.1.Self-supervised graph-masked modeling

Pretraining masks a random fraction ρ = 0.15 of observed biomarkers and reconstructs the hidden values from the remaining graph context. For masked index set Mi, the reconstruction loss is

$$\mathcal{L}_{\mathrm{GMM}} = \frac{1}{\sum_i |\mathcal{M}_i|} \sum_i \sum_{j \in \mathcal{M}_i} \left(\widehat{b}_{ij} - b^{\star}_{ij}\right)^2 .$$

This objective learns biomarker embeddings without using weak disease la-bels. Site-wise pretraining is simulated through FedAvg-style aggregation across environment-defined clients. For client c, update Δc is clipped and noise-perturbed before aggregation:

$$\tilde{\Delta}_c = \Delta_c \cdot \min\left(1, \frac{C}{\|\Delta_c\|_2}\right) + \mathcal{N}(0, \sigma^2 C^2 I), \qquad w^{(t+1)} = w^{(t)} + \eta \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \tilde{\Delta}_c. \tag{6}$$

The study uses this clipping-plus-noise loop as a privacy-aware simulation inspired by differential privacy and FedAvg [37, 43]; it does not claim formal (ε, δ)-differential privacy or secure aggregation.

After pretraining, shallow supervised heads predict multi-label risk and discrete-time survival. The risk head uses binary cross-entropy over three weak labels. For survival, each observed time ti maps to bin

$$\beta_i = \min\left\{K_s, \left\lfloor \frac{t_i}{t_{\max}} (K_s - 1) \right\rfloor + 1\right\}.$$

The per-sample survival log-likelihood is

$$\ell_i^{\mathrm{surv}} = \sum_{j=1}^{\beta_i - 1} \log(1 - h_{ij}) + \begin{cases} \log h_{i\beta_i}, & \delta_i = 1, \\ \log(1 - h_{i\beta_i}), & \delta_i = 0, \end{cases} \qquad \mathcal{L}_{\mathrm{surv}} = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \ell_i^{\mathrm{surv}}. \tag{7}$$

To reduce environment-specific shortcuts, the supervised objective adds an anchor penalty on environment-aligned residuals. Let A denote one-hot environment anchors in a mini-batch, PA = A(AᵀA)⁻¹Aᵀ, and R = Y − Y . The nuisance-free anchor penalty is

$$\mathcal{R}_{\mathrm{anc}} = \frac{1}{B} \|P_A R\|_F^2 .$$

The full training objective is

$$\mathcal{L}_{\mathrm{joint}} = \mathcal{L}_{\mathrm{risk}} + \mathcal{L}_{\mathrm{surv}} + \lambda_{\mathrm{anc}} \mathcal{R}_{\mathrm{anc}} + \lambda_{\mathrm{phys}} (\mathcal{R}_{\mathrm{GI}} + \mathcal{R}_{\mathrm{KD}} + \mathcal{R}_{\mathrm{BP}}). \tag{8}$$

The main configuration uses AdamW with learning rate 2 × 10−3, supervised gradient clipping at g 2 ≤ 2, λanc = 0.1, Ks = 20 survival bins, ReduceL-ROnPlateau on calibration BCE, and early stopping with patience 2. These settings keep the model compact while maintaining a clear separation be-tween training, calibration, and test evaluation.

Algorithm 1 Per-label split conformal thresholding for multi-label risk

Require: Calibration probabilities p_{ik}, labels y_{ik}, target level α

1: **for** $k = 1, \ldots, 3$ **do**
2: Compute $s_{ik} = 1 - p_{ik}$ if $y_{ik} = 1$, else $s_{ik} = p_{ik}$
3: Set $\hat{q}_{1-\alpha,k}$ to the finite-sample $(1 - \alpha)$ quantile of $\{s_{ik}\}$
4: Set $\tau_k = 1 - \hat{q}_{1-\alpha,k}$
5: **end for**
6: Return $\hat{Y}(x) = \{k : p_k(x) \geq \tau_k\}$

3.4.Calibration, Robustness, Fairness, and Governance

3.4.1.Per-label conformal thresholding

CIPHER–Twin converts risk probabilities into set-valued predictions us-ing label-wise split conformal calibration [28, 29, 30]. For calibration sample i and label k, with predicted probability pik, the nonconformity score is

$$s_{ik} = \begin{cases} 1 - p_{ik}, & y_{ik} = 1, \\ p_{ik}, & y_{ik} = 0, \end{cases} \qquad \text{with target coverage } 1 - \alpha. \tag{9}$$

The quantile q1−α,k of calibration scores defines the threshold τk = 1 − q1−α,k,

and the predi^ction set is

$$\widehat{\mathcal{Y}}(x) = \{k : p_k(x) \geq \tau_k\}. \tag{10}$$

Label-wise calibration is used because the three weak labels come from dis-tinct source distributions and have strong imbalance.

3.4.2.Drift monitoring

At deployment time, nonconformity scores can be monitored through an exponentially weighted moving average:

$$z_t = \lambda r_t + (1 - \lambda) z_{t-1}, \qquad z_0 = \mathrm{median}\{r_i : i \in \mathcal{I}_{\mathrm{cal}}\}. \tag{11}$$

Here, rt denotes the incoming nonconformity score, and λ controls smoothing. An alert is triggered when zt exceeds a calibration-derived control threshold, such as the 95th percentile of calibration EWMA values. This monitor provides a simple warning for covariate shift or label shift without changing model weights.

3.4.3.Group-wise calibration and fairness

Fairness analysis is performed only for groups available in the public co-horts, mainly sex-defined groups. The evaluation reports group AUC gaps,

$$\Delta\mathrm{AUC}_k = \max_g \mathrm{AUC}_{k,g} - \min_g \mathrm{AUC}_{k,g},$$

and conformal positive-class coverage gaps,

$$\Delta\mathrm{Cov}_k^{+} = \max_g \mathrm{Cov}_{k,g}^{+} - \min_g \mathrm{Cov}_{k,g}^{+}.$$

These metrics check whether discrimination and coverage remain stable across groups. The analysis does not make any claim to clinical fairness as the public datasets included limited social and structural variables.

All experiments are based on public, de-identified data and there is no attempt to link data across cohorts. The framework is designed for methodological research and engineering evaluation, but not at the bedside. The audit package contains all the schema maps, ontology tables, missingness summaries, seeds, software versions, vector figures, and exported metric tables. Weak labels are treated as cohort-provenance proxies and don't make causal or patient-level clinical claims from these labels. For any planned deployment to include patient level deduplication, a re-view by the local IRB, formal privacy accounting, secure aggregation, subgroup bias audits, shadow testing, and human escalation in case that the model abstains or detects drift.

4. EXPERIMENTAL SETUP AND BASELINES

Before presenting the empirical results, this section describes the evalua-tion tasks, data splits, metrics, baseline models, and implementation details used to assess CIPHER–Twin.

4.1.Tasks and Data Splits

The empirical study covers two families of tasks defined in Sec. 3: (i) multi-label chronic disease risk prediction and (ii) discrete-time survival anal-ysis.

Multi-label risk prediction.. The first task predicts three binary labels of broad clinical interest for each record in the heterogeneous union: cardiovas-cular disease (CVD), type 2 diabetes mellitus (T2DM), and chronic kidney disease (CKD). The label vector Y (x) = (YCVD, YT2DM, YCKD) ∈ {0, 1}3 fol-lows the weak label convention in Eq. (3), with proxy positives tied to cohort provenance (Framingham for CVD, Diabetes 130–US for T2DM, CKD cohort for CKD). All models, including baselines, operate on the same harmonized tabular inputs (Sec. 3) and predict these three labels jointly or independently.

Discrete-time survival.. The second task models time-to-death in the Heart Failure Clinical Records cohort using the discrete-hazard parameterization in Eq. (4). Each subject has follow-up time Ti (days) and event indicator δi ∈
{0, 1}. Hazards hk(x) ∈ (0, 1) over K time bins induce a survival curve Sk(x), which supports standard concordance-based evaluation and calibration.

Row-level partitioning.. To separate model fitting, calibration, and evalua-tion, the heterogeneous union uses a fixed row-level split into training, cal-ibration, and test sets in the ratio 60%:20%:20%. The split is stratified by environment tag E ∈ {framingham, diabetes, ckd, heartfail} and by the three proxy labels, so that each subset contains examples from all cohorts and labels where possible. The calibration subset is used only for:

- per-label conformal thresholding and set-valued risk construction (Sec. 3.4.1),
- learning-rate scheduling and early stopping,

while the test subset remains untouched until final evaluation. The Di-abetes cohort contains multiple encounters per patient identifier; the split therefore acts as an end-to-end stress test of robustness rather than a strictly deduplicated benchmark, and this limitation is stated explicitly in the inter-pretation of results.

Site-wise survival split.. For the survival task on Heart Failure records, the
same global split is used. All subjects from this cohort inherit the train/calibration/test assignment from the union, and only those subjects contribute to the survival head loss in Eq. (7). This choice keeps the multi-task setting consistent and avoids a separate survival-only partition.

4.2.Evaluation Metrics

The evaluation uses standard metrics for discrimination, calibration, sur-vival performance, and fairness, so that results remain comparable to prior work on EHR risk prediction and survival modelling.

Discrimination (classification).. For each of the three proxy labels (CVD, T2DM, CKD), discrimination on the test split is summarized by:

- Area under the receiver operating characteristic curve (AUROC).
- Area under the precision–recall curve (AUPR).
- F1-score at an operating point chosen by maximizing F1 on the cali-bration split.

Let pˆik denote the predicted probability for label k on subject i, and
yik ∈ {0, 1}. AUROC and AUPR summarize ranking quality; F1 captures a single-threshold balance between precision and recall.

Calibration (classification).. Probability calibration is assessed in two ways:

- Negative log-likelihood (NLL) and Brier score

$$\text{Brier}_k = \frac{1}{N}\sum_{i=1}^{N} \hat{p}_{ik} - y_{ik}^{2},$$

computed per label on the test split.

- Reliability diagrams and expected calibration error (ECE) with equal-width probability bins, both overall and by subgroup when sex is avail-able (Secs. 3.4.1 and 3.4.3).

These metrics quantify how well predicted probabilities reflect empirical event frequencies before conformalization.

Conformal coverage and set size.. After learning per-label thresholds on the calibration split (Algorithm 1), the test split evaluation reports:

- positive-class coverage Cov+ = P(k ∈ Y(X) | Yk = 1),
- empirical distribution of set sizes |Y(x)|,

for each label k. These values measure how often the true label lies inside the predicted set and how large those sets are in practice.

Survival metrics.. For the discrete-time survival head, the evaluation reports:

- Harrell's concordance index (C-index) on the test subset of Heart Fail-ure patients, which measures the probability that predicted risks cor-rectly order observed times.
- Integrated Brier score (IBS) over the time range, computed with right-censoring handled by inverse probability of censoring weights.

These metrics summarize discrimination and calibration of survival curves, respectively.

Fairness and robustness.. To assess robustness across subgroups and envi-ronments, the analysis computes:

- group-wise AUROC per label and the AUROC gap ΔAUCk between groups defined by sex when available,
- group-wise positive-class conformal coverage and the coverage gap ΔCov+

(Sec. 3.4.3),

- EWMA-based drift alerts over a stream of nonconformity scores, as described in Sec. 3.4.2.

These quantities indicate whether performance and coverage remain sta-ble across groups and under moderate distribution shift.

4.3.Baseline Models

CIPHER–Twin is compared against a set of classical and neural base-lines that operate on the same harmonized tabular inputs and use the same train/calibration/test split.

Tabular risk baselines.. For multi-label risk prediction, the following models act as baselines:

- Logistic regression (LR). Independent ℓ2-regularized logistic models for each of the three labels, trained on the harmonized numeric features (with simple mean imputation and standardization inside each training set). Regularization strength is chosen by five-fold cross-validation on the training subset.
- Random forest (RF). A multi-label random forest with class-weighted impurity, applied to the same features as LR. The number of trees and maximum depth are tuned on the calibration split.
- Gradient boosted trees (GBT). A tree-based gradient boosting model for each label using shallow trees and learning-rate shrinkage. Hyperparameters (number of estimators, depth, learning rate) are tuned on the calibration split.
- Shallow multilayer perceptron (MLP). A fully connected network with one hidden layer, ReLU activations, and dropout, trained with AdamW on the same train/calibration split as CIPHER–Twin. This baseline tests whether a simple neural tabular model can match the proposed encoder without graph structure or invariance objectives.

All baselines receive exactly the same harmonized predictors as the Sim-pleTwinEncoder (Sec. 3.2), but do not use typed graphs, anchor-based in-variance, or conformal calibration. For a fair comparison, each baseline uses early stopping based on calibration-split NLL and adopts the same evaluation metrics as the proposed model.

Survival baselines.. For the survival task on Heart Failure records, two stan-dard models form the comparison set:

- Cox proportional hazards (CoxPH). A semi-parametric Cox model fitted with ℓ2 regularization. Continuous predictors are standardized; categorical predictors are one-hot encoded. The concordance index and IBS are computed on the same test subset as CIPHER–Twin.
- Random survival forest (RSF). An ensemble of survival trees with log-rank splitting, tuned by number of trees and node size on the cal-ibration split. RSF naturally handles non-linearities and interactions in the baseline feature space.

These baselines represent widely used survival modelling approaches in clinical risk prediction and serve as reference points for the discrete-hazard head in CIPHER–Twin.

Ablated versions of CIPHER–Twin.. In addition to external baselines, sev-eral ablated variants of CIPHER–Twin are evaluated to isolate the effect of architectural choices:

- No-anchor variant. Removes the anchor-based invariance penalty

Ranc while keeping all other components fixed.

- No-SSL variant. Trains the supervised heads from scratch without graph-masked pretraining (Sec. 3.3.1).
- SDoH-only student. Uses only the distilled SDoH encoder (age, sex, smoking) without biomarker embeddings, reflecting a low-resource triage scenario.

These ablations help attribute gains to invariance, self-supervision, and richer inputs rather than to generic neural capacity.

4.4.Implementation Details and Reproducibility

All models share a common implementation and evaluation pipeline to support reproducibility.

Software stack.. The reference implementation uses Python with PyTorch for neural components, together with standard libraries for data handling and classical baselines. Gradient-based models use the AdamW optimizer. All random-number generators (Python, NumPy, PyTorch) receive fixed seeds, and deterministic flags are enabled where available to reduce run-to-run vari-ance.

Training protocol.. For CIPHER–Twin, the training procedure follows Sec. 3.3. Self-supervised graph-masked pretraining runs separately on each environ-ment with encoder parameters shared via a small number of FedAvg rounds and DP-style clipping and noise. After pretraining, the multi-label risk head and discrete-hazard survival head are trained jointly with the joint loss in Eq. (8). Early stopping monitors calibration-split NLL, and a Reduce-on-Plateau scheduler adapts the learning rate. For baselines, training uses early stopping on calibration-split NLL or partial log-likelihood (for CoxPH).

Hardware.. The encoder and neural baselines run on a single modern GPU when available, but the model size and batch configuration also permit CPU-only execution for smaller experiments. Gradient clipping keeps training numerically stable, and memory usage scales linearly with batch size and number of observed biomarkers.

Table 2: Cohort scale and missingness after harmonization. Overall missingness is the mean fraction of NaN values across all cells.

Dataset	Rows	Vars	Overall missing (%)	Main missingness pattern
Framingham	4,240	16	0.95	Low missingness; main gap in glucose_random
				(9.15%); chol_total missing in 1.18%.
Diabetes	101,766	50	7.35	Strong block missingness in weight (96.86%),
				max_glu_serum (94.75%), and A1Cresult
				(83.28%); therapy and discharge fields are
				complete.
CKD	400	26	9.70	Moderate missingness; largest gaps in rbc

				(38.00%), rbc_count (32.50%), wbc_count
				(26.25%), sod (21.75%), and pot (22.00%).
Heart Failure	299	13	0.00	Complete harmonized variables; small cohort size
				limits survival modelling strength.

5. RESULTS

5.1.Cohort Profiles and Missingness

Table 2 and Figure 1 summarize cohort scale and missingness after har-monization. The four cohorts differ strongly in sample size, feature availabil-ity, and missingness structure. Heart Failure is complete but small, Fram-ingham has low overall missingness, Diabetes has block missingness in lab-oratory proxy fields, and CKD shows moderate missingness in hematology, urine, and electrolyte variables. This pattern supports the use of NaN-safe encoding instead of early global imputation.

Figure 2 shows two shift diagnostics. The proxy-risk distribution differs across environments, and the learned embedding space forms large cohort-aligned clusters with local overlap. These results confirm that cohort identity remains a strong source of variation, which justifies the environment anchor and conformal calibration steps used in the framework.

5.2.Discrimination and Calibration

Table 3 and Figure 3 report test-set discrimination for the three weak-label risk tasks. CVD and T2DM reach near-ceiling AUROC values, while CKD remains high in AUROC but has lower AUPR. The lower CKD AUPR reflects lower prevalence and weaker renal biomarker observability. These val-ues should be interpreted as performance under environment-as-label super-vision, not as prospective incident-risk performance in a patient-deduplicated clinical cohort. Figure 4 summarizes reliability. T2DM probabilities concentrate near 0 or 1, which matches the near-deterministic separation in Figure 3. CKD

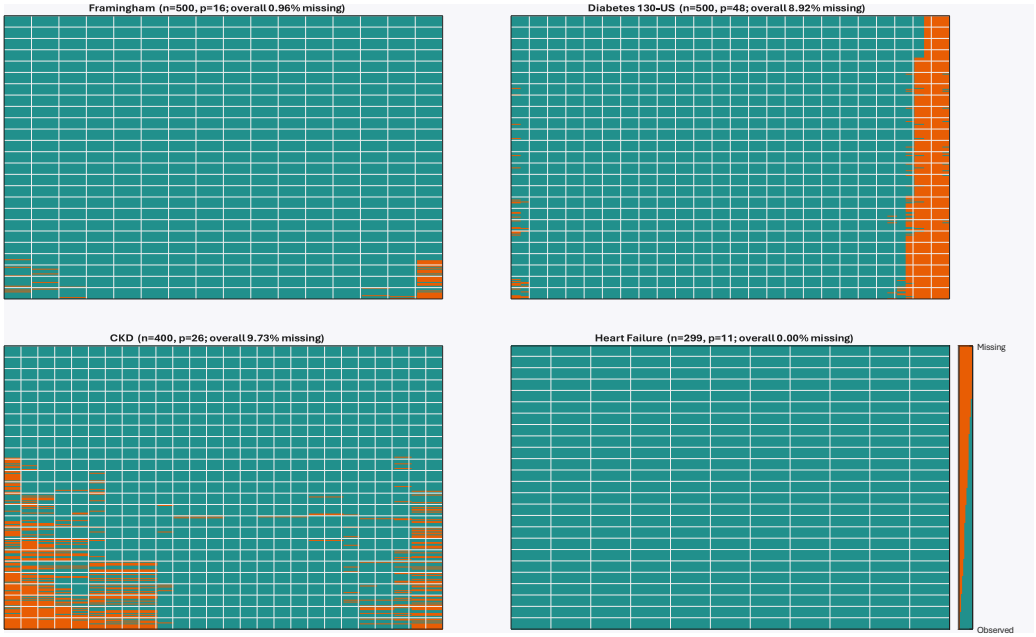
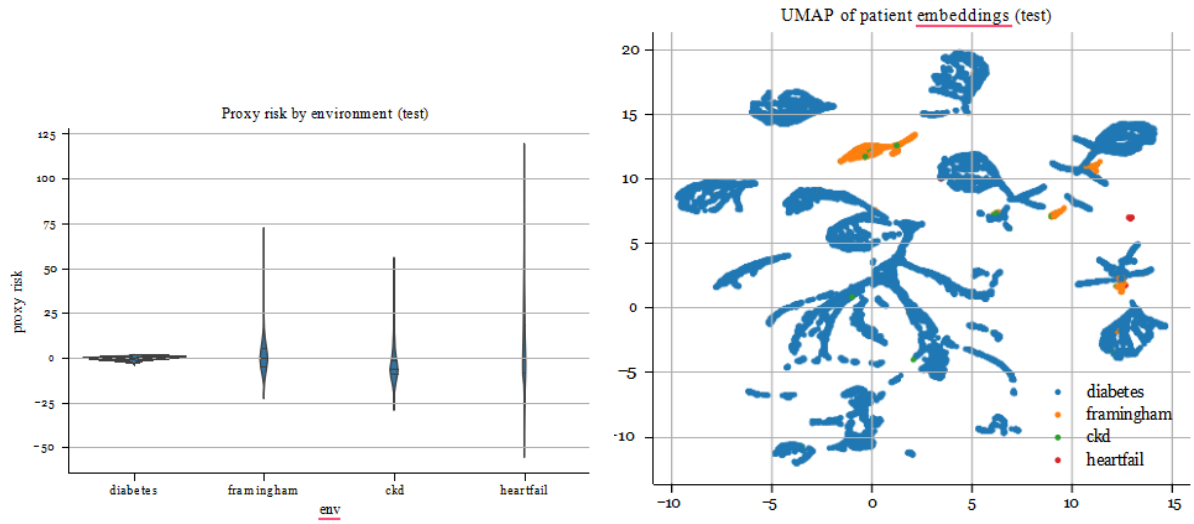


Figure 1: Record-wise missingness heatmaps for the four cohorts using the first 500 rows per cohort. Diabetes shows block missingness in laboratory proxies, CKD shows sparse laboratory fields, Framingham is nearly complete, and Heart Failure is complete.



Proxy risk by environment **Embeddings by environment**

Figure 2: Environment shift diagnostics on the test split. Left: proxy-risk score differs across cohorts. Right: UMAP of learned patient embeddings shows broad cohort-aligned structure with some local overlap.

Table 3: Bootstrap discrimination on the test split with 1000 resamples.

Label	AUROC	AUPR
CVD	0.9986 [0.9981–0.9991]	0.9216 [0.8923–0.9506]
T2DM	0.9998 [0.9995–1.0000]	1.0000 [0.99998–1.00000]
CKD	0.9732 [0.9653–0.9794]	0.3313 [0.2334–0.4425]

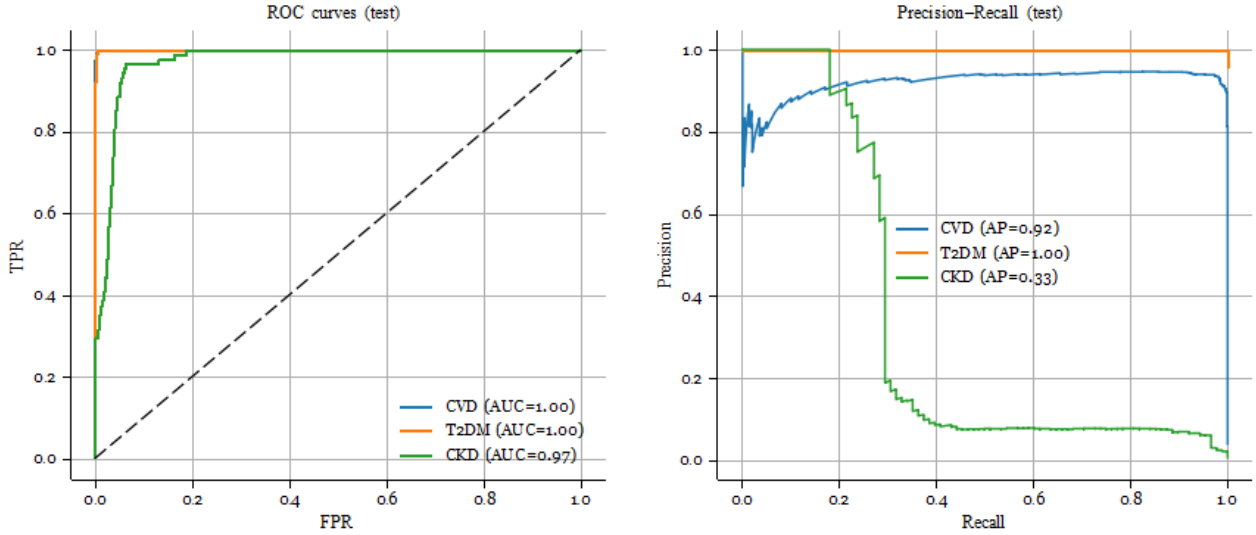


Figure 3: Test-set discrimination. Left: ROC curves. Right: precision–recall curves. CVD and T2DM show strong separation, while CKD has lower precision–recall performance under class imbalance and sparse renal observability.

overpredicts in some mid-range bins, consistent with sparse renal signals. CVD calibration remains close to the identity line in the pooled and sex-stratified plots, with no gross sex-specific miscalibration visible in this test split.

5.3.Conformal Coverage and Set Size

Table 4 reports per-label thresholds and positive-class coverage. The thresholds follow task difficulty: T2DM receives a very high threshold be-cause its scores are decisive, CKD needs a very low threshold to maintain coverage under weaker separation, and CVD falls between these two cases. Test coverage stays close to the 90% target for all three labels.

The mean prediction-set size is approximately 0.95, and most outputs are single-label sets. Empty sets appear in a small fraction of cases and serve as abstentions. Figure 6 shows the expected coverage–size trade-off: as α increases, coverage decreases and mean set size shrinks.

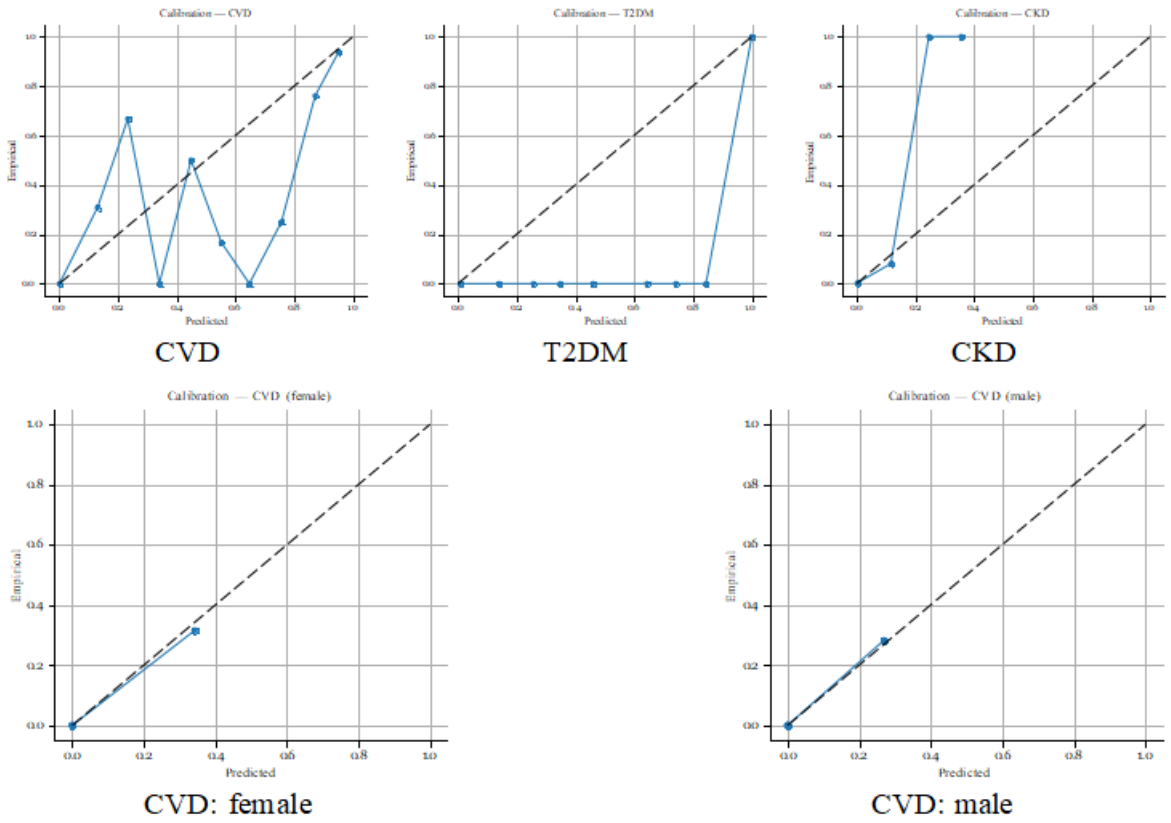


Figure 4: Reliability plots on the test split. The top row shows pooled calibration for CVD, T2DM, and CKD. The bottom row shows CVD calibration by sex.

Table 4: Per-label conformal thresholds and test-set positive-class coverage. Sex-stratified values are shown where positive cases exist.

CVD	**0.9253**	**0.878**	**(0.914, 0.850)**
T2DM	0.999965	0.902	(0.930, 0.878)
CKD	0.00352	0.920	(n/a, 0.920)

5.4.Survival Analysis

The Heart Failure cohort provides 299 patients, 96 deaths, and a median follow-up of approximately 115 days. Figure 7 shows that predicted median-survival bins are not collapsed into a single bin, which indicates coarse sep-aration of early, intermediate, and late horizons. However, Table 5 shows a C-index of 0.419, so the survival head does not provide reliable individual-level ranking in this static and small cohort.

This result limits the survival claim to feasibility only. The main con-

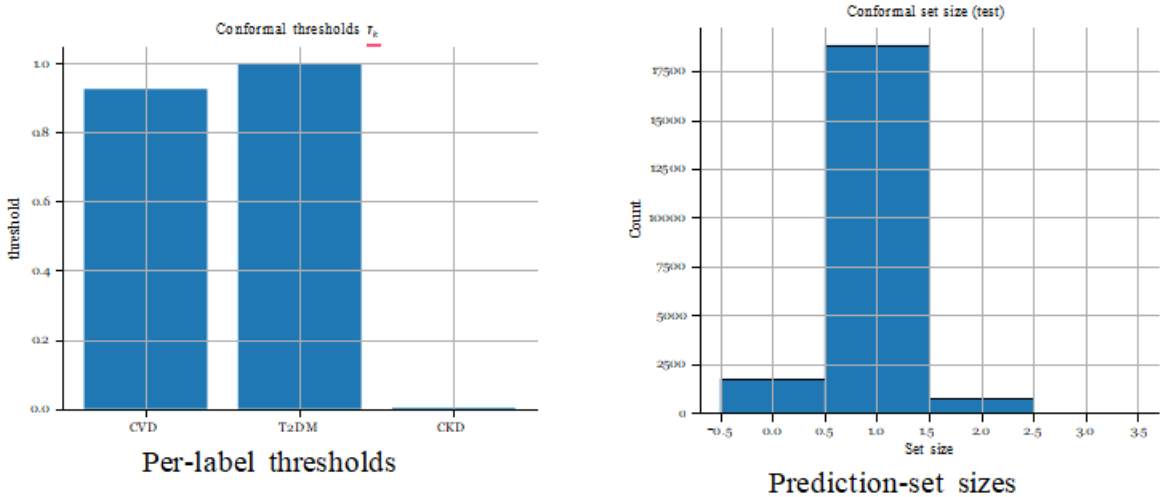


Figure 5: Conformal outputs on the test split. Left: calibrated thresholds. Right: most prediction sets are singletons, while empty sets act as abstentions for low-signal cases.

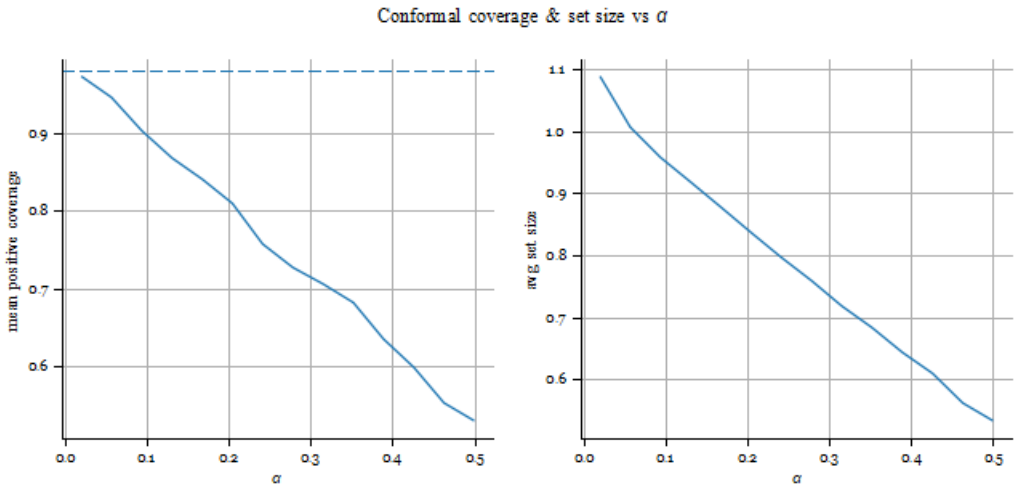


Figure 6: Coverage and mean set size as functions of the nominal error level α. Higher reliability requires larger prediction sets, while higher α yields smaller and more decisive sets.

Table 5: Time-to-event performance on the Heart Failure cohort.

Cohort	Patients / Events	Follow-up median	C-index
Heart Failure	299 / 96	≈ 115 days	0.419

straint is the small, static heart-failure file with no longitudinal laboratory sequence. Richer time-stamped covariates and survival-specific calibration are needed before using the survival head for clinical prognosis.

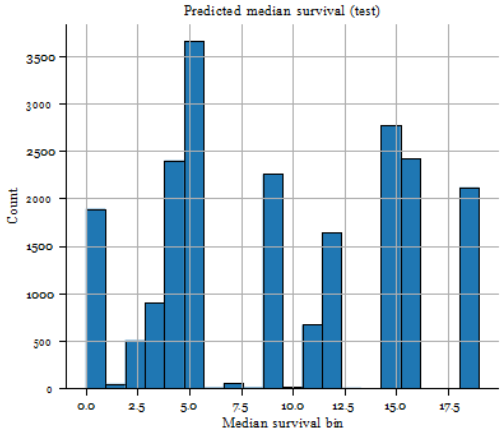


Figure 7: Predicted median-survival bins on the test split. The multi-modal distribution suggests coarse horizon separation, but ranking quality remains weak.

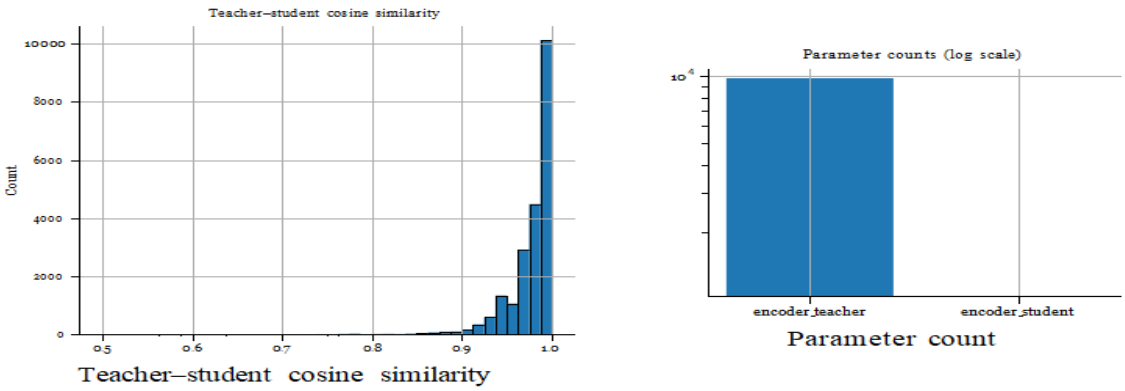


Figure 8: Distillation and efficiency results. The student embedding aligns closely with the teacher embedding while reducing model size substantially.

5.5.Efficiency and Distillation

Figure 8 and Table 6 show that the SDoH-only student closely matches the teacher embedding direction while using far fewer parameters. The cosine similarity distribution peaks near 1.0, and the student reduces the encoder size from 9,809 to 1,152 parameters, giving an approximately 8.5× parameter reduction.

Table 6: Model complexity and approximate memory footprint using 32-bit weights.

Model	Params	log10(Params)	Footprint (KB)	Reduction
Teacher encoder	9,809	3.99	≈ 38	–
Student encoder (SDoH)	1,152	3.06	≈ 4.5	≈ 8.5× fewer params

These results support a two-tier inference design: the student can provide fast SDoH-only embeddings in low-resource settings, while the full encoder remains preferable when laboratory biomarkers are available. Embedding alignment does not guarantee identical calibrated probabilities, so the student still requires task-specific recalibration before independent use.

6. PROTOTYPE GUI AND DEPLOYMENT

A lightweight graphical interface accompanies the research codebase and is intended for local experimentation rather than public hosting.1 The GUI mirrors the harmonized schema and the typed patient–phenotype graph, ren-dering inputs and outputs in a way that foregrounds uncertainty and absten-tion. All results shown here were produced offline using this interface; no patient data were transmitted to external services.

6.1.GUI Features and Inference Pipeline

The single–case screen (Figure 9) exposes SDoH and biomarker fields that survived harmonization (Sec. 3). Users may leave fields blank; the en-coder's nan–safe operators treat missingness as signal rather than imputing across records. On submit, the interface executes the same inference path as in training: robust within–environment normalization; construction of a pa-tient–centric subgraph; encoding to $z \in R^d$ via SimpleTwinEncoder; evalua-tion of the risk heads for {CVD,T2DM,CKD}; and optional discrete–hazard survival outputs. Probabilities are then mapped to set–valued predictions us-ing fixed conformal thresholds $\{\tau_k\}$ learned on the calibration split; the dis-play emphasizes the set decision (in/out) and provides the target coverage $1 - \alpha$, in line with distribution–free guarantees [28, 30]. Because raw con-fidence can be misread by end–users, the GUI visually down–weights small probability differences and highlights abstentions [4].

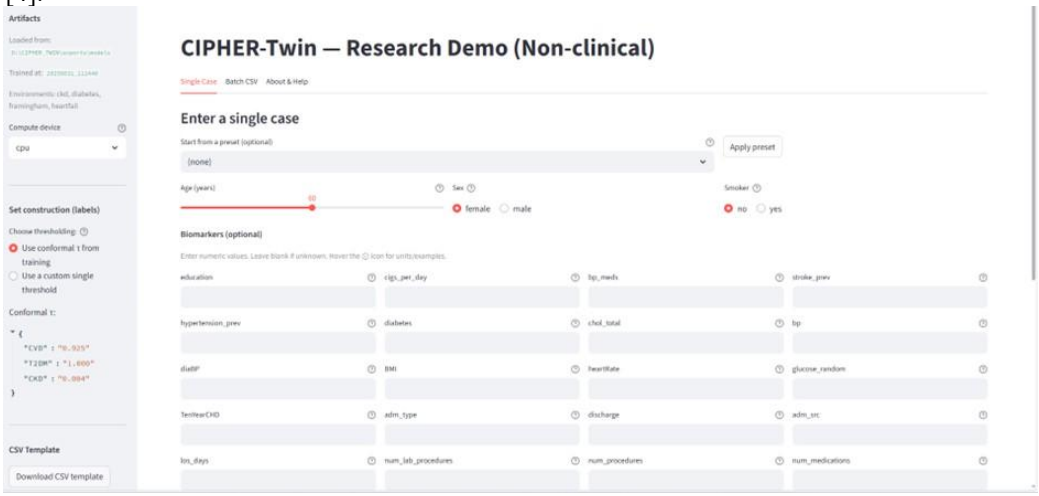


Figure 9: Single–case interface (local run). Inputs reflect the harmonized schema; outputs report calibrated probabilities and set–membership under conformal thresholds τ_k.

6.2.Batch CSV Scoring and Exports

A batch mode accepts a template CSV consistent with the harmonized ontology and streams rows through the same inference pipeline. The re-turned table contains per–label probabilities, conformal set membership at the chosen $1 - \alpha$, and, when enabled, survival summaries (median survival bin and per–bin hazards). Figure 10 illustrates the batch screen and an example output where in–set indicators correspond to the fixed $\{\tau_k\}$ from calibration. The compact encoders support responsive CPU inference in this offline setting.

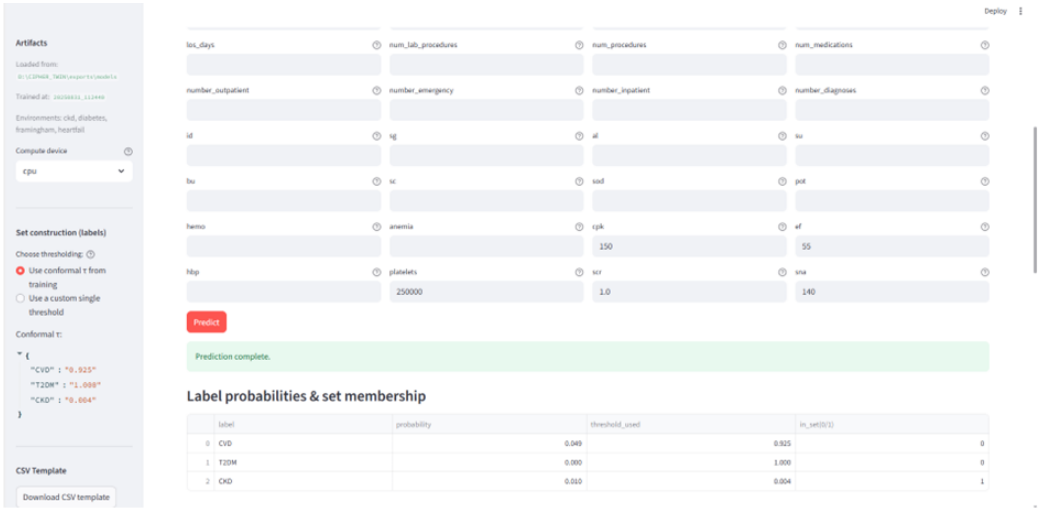


Figure 10: Batch CSV mode (local). Rows are scored with the same normalization, graph construction, and conformal thresholding as in training; outputs enable cohort–level auditing of calibration and coverage.

6.3.Reference Deployment and Security Considerations

Although the present study runs the GUI locally, prospective clinical integrations would require a hardened service. Table 7 summarizes recom-mended controls for a minimal cloud deployment. These measures reduce risk but do not replace site governance, IRB review, or formal privacy ac-counting. In particular, if federated training were pursued, client updates should be per–example clipped and perturbed with noise tracked by a pri-vacy accountant before any differential–privacy claim [43]. We also recom-mend inference–only public endpoints and on–prem execution of the distilled student model when compute or data–sharing constraints apply.

Table 7: Recommended controls for a minimal cloud deployment

Domain	Control	Residual risk
Network isolation	Private subnet; TLS; restricted ingress; admin endpoints behind allow–lists	Misconfiguration; credential leakage
AuthZ/AuthN	Role–based access; short–lived tokens; clear separation of score vs. view roles	Session hijack if client compromised
Data handling	Encrypted volumes; ephemeral job storage; no payloads in logs	Forensic recovery if host compromised
Model integrity	Versioned, checksummed model&threshold bundles; controlled rollbacks	Drift if thresholds not refreshed with recalibration
Observability	De–identified metrics (latency, error codes) for monitoring; alerting on drift	Rare failure modes may escape detection
Privacy posture	Inference–only public service; no training endpoints; formal DP accounting if FL used	Membership–inference risk if parameters leak

7. CONCLUSION

In practice, the EHR data are partially observed, heterogeneous tables with varying sets of features, coding conventions, prevalence, and follow-up. Risk models novel engineering-centric digital twin system, CIPHER–Twin, which mitigates the above limitation by introducing typed patient–phenotype graphs, a compact tabular encoder, anchor-based invariance, conformal prediction, and a federated-ready train-ing loop.

should be able to cope with dataset shift, weak and delayed la-bels, and stringent calibration, fairness and privacy requirements. Based on public EHR-style cohorts, this study presented a

The design represents each record as a typed graph of connections between patients, social determinants of health, organs, biomarkers and therapies, and maps this graph to a latent representation using the SimpleTwinEncoder. This encoder is able to learn stable embeddings of biomarkers, conditional on social

determinants, without early imputation or leakage between samples, thanks to graph-masked self-supervision and site-wise robust scaling. Super-imposed on this latent space, there is a multi-label head for predicting chronic disease risk for cardiovascular disease, type 2 diabetes, and chronic kidney disease; and a discrete-hazard survival head for modelling time-to-event out-comes for heart failure. On the four public cohorts, this combination shows very good discrimination for CVD and T2DM risk, good discrimination for CKD, and near-nominal coverage when thresholds are used to transform scores into set valued prediction for the limited heart-failure survival data, and competitive discrimination on these cohorts.

The primary emphasis of the framework is on robustness and reliability under cohort shift. Anchor-based invariance introduces the notion of ob-servable anchors, the environment tags, and penalizes environment-aligned residuals, thus decreasing the need to take shortcuts for a particular site and making performance more stable across datasets. Finally, per-label split conformal prediction can be used to transform these raw probabilities into calibrated sets of finite size with close-to-target coverage guarantees that con-tain the true label. An EWMA monitor on nonconformity scores provides a simple way to monitor for drift at deployment. These elements enable a standard neural tabular encoder to become a digital twin, exposing its uncertainty while responding to distribution shift in an auditable way.

The privacy and deployability are explicitly considered, with a federated-ready training strategy and a distilled student encoder. Federated simulation is based on FedAvg, norm clipping and additive Gaussian noise on client up-dates to investigate the behavior of the encoder when the training data are kept in simulated training sites. While the present work does not explicitly provide differential privacy guarantees or secure aggregation, the clipping-plus-noise pipeline is consistent with typical DP pipelines and provides a glimpse into where a production system would require more privacy account-ing and secure infrastructure. The distilled SDoH-only student compresses the entire encoder into a small model that needs only age, sex and smoking status and maintains a significant proportion of the triage or edge ranking capabilities of the teacher, showing a route towards low latency triage or edge deployment in resource constrained areas.

The conclusions are limited in a number of ways. The multi-label task is labeled with an environment-as-label approach, meaning that labels are still weak proxies of cohorts of provenance. The Heart Failure cohort is small and shallow in time, meaning that the survival head is not in a definite clinical benchmark, but in a stress-test regime. The federated learning component is an offline simulation with no actual institutional boundaries, formal pri-vacy budgets or secure aggregation, and as such cannot make strict privacy claims. Group-wise fairness analysis only includes sex as reported in the public datasets and does not capture the breadth and depth of social and structural factors that contribute to inequities in care. Physiological priors for glucose, kidney function and blood pressure are included as diagnostics and for the most part are not active in the main experiments due to the virtually cross-sectional nature of the cohorts.

Nevertheless, CIPHER–Twin serves as a pragmatic set of instructions on how to create more robust and transparent EHR models in the face of realistic constraints. Typed patient–phenotype graphs are a simple but pow-erful way to standardize inputs from various data tables. Models can deal with heavy missingness and weak supervision with mask-aware encoders and graph-masked self-supervision. Combined, anchor-based invariance and con-formal prediction address dataset shift and uncertainty and provide metrics that can be understood by clinicians, including coverage and abstention, and not just AUROC. A public-cohort experiment opens the door to privacy-aware, resource-aware deployment via federated-ready training and a distilled student.

This framework can be extended in several ways in the future. The cur-rent diagnostic penalties could be replaced with active regularizers on real trajectories, using richer longitudinal EHRs, imaging data, and/or physiologi-cal waveforms to support the use of physiological priors and neural differential equations. In particular, simulated federated loop might be replaced with a real cross-site deployment, including secure aggregation, formal moments or Rényi privacy accounting, and institution-specific governance, in the form of multi-institution collaborations. Better causal interpretation and more accu-rate labels would be enabled by patient-level deduplication and more accurate labeling. Finally, group-conditional conformal methods and fairness-aware objectives might aim at achieving equitable coverage for each group defined by sex, age, and other approved characteristics. Together, these extensions would bring CIPHER–Twin closer to being a deployable, clinical data, regu-latory and operational-aware digital twin for chronic disease management.

Declarations

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Suman: Conceptualization, Methodology, Software, Data curation, Writ-ing—original draft. Yudhvir Singh, & Neha Gulati: Investigation, Su-pervision, Validation, Writing—review & editing.

Funding

No funding was received for the study.

Data Availability

The datasets used during the current study available from the correspond-ing author on reasonable request.

REFERENCES

1. B. Shickel, P. J. Tighe, A. Bihorac, P. Rashidi, Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis, IEEE journal of biomedical and health informatics 22 (5) (2017) 1589–1604.
2. R. Miotto, L. Li, B. A. Kidd, J. T. Dudley, Deep patient: an unsuper-vised representation to predict the future of patients from the electronic health records, Scientific reports 6 (1) (2016) 26094.
3. J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence, Dataset Shift in Machine Learning, The MIT Press, 2009.

4. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of mod-ern neural networks, in: International conference on machine learning, PMLR, 2017, pp. 1321–1330.
5. R. N. Bergman, L. S. Phillips, C. Cobelli, et al., Physiologic evalua-tion of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose., The Journal of clinical investigation 68 (6) (1981) 1456–1467.
6. L. Magni, D. M. Raimondo, C. Dalla Man, G. De Nicolao, B. Kovatchev, C. Cobelli, Model predictive control of glucose concentration in type i diabetic patients: An in silico trial, Biomedical Signal Processing and Control 4 (4) (2009) 338–346.
7. A. C. Guyton, J. E. Hall, Guyton and Hall textbook of medical physi-ology, Elsevier, 2011.
8. S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, P. E. Johnson, P. J. O'Connor, Data mining for censored time-to-event data: a bayesian network model for predict-ing cardiovascular risk from electronic health record data, Data Mining and Knowledge Discovery 29 (4) (2015) 1033–1069.
9. J. Wolfson, S. Bandyopadhyay, M. Elidrisi, G. Vazquez-Benitez, D. M. Vock, D. Musgrove, G. Adomavicius, P. E. Johnson, P. J. O'Connor, A
10. naive bayes machine learning approach to risk prediction using censored, time-to-event data, Statistics in medicine 34 (21) (2015) 2941–2957.
11. B. Liu, Y. Li, Z. Sun, S. Ghosh, K. Ng, Early prediction of diabetes com-plications from electronic health records: A multi-task survival analysis approach, in: Proceedings of the AAAI Conference on Artificial Intelli-gence, Vol. 32, 2018.
12. R. Sharma, H. Anand, Y. Badr, R. G. Qiu, Time-to-event predic-tion using survival analysis methods for alzheimer's disease progression, Alzheimer's & Dementia: Translational Research & Clinical Interven-tions 7 (1) (2021) e12229.
13. A. B. Teshale, H. L. Htun, M. Vered, A. J. Owen, R. Freak-Poli, A sys-tematic review of artificial intelligence models for time-to-event outcome applied in cardiovascular disease risk prediction, Journal of medical sys-tems 48 (1) (2024) 68.
14. X. Yang, H. Qiu, L. Wang, X. Wang, Predicting colorectal cancer sur-vival using time-to-event machine learning: retrospective cohort study, Journal of Medical Internet Research 25 (2023) e44417.
15. M. Zisser, D. Aran, Transformer-based time-to-event prediction for chronic kidney disease deterioration, Journal of the American Medical Informatics Association 31 (4) (2024) 980–990.
16. X. Liu, H. Wang, T. He, Y. Liao, C. Jian, Recent advances in represen-tation learning for electronic health records: A systematic review, in: Journal of Physics: Conference Series, Vol. 2188, IOP Publishing, 2022, p. 012007.
17. E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, J. Sun, Gram: graph-based attention model for healthcare representation learning, in: Pro-ceedings of the 23rd ACM SIGKDD international conference on knowl-edge discovery and data mining, 2017, pp. 787–795.
18. E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, Advances in neural information processing systems 29 (2016).
19. S. G. Paul, A. Saha, M. Z. Hasan, S. R. H. Noori, A. Moustafa, A sys-tematic review of graph neural network in healthcare-based applications: Recent advances, trends, and future directions, IEEE Access 12 (2024) 15145–15170.
20. T. Tang, Z. Han, Z. Cai, S. Yu, X. Zhou, T. Oseni, S. K. Das, Personal-ized federated graph learning on non-iid electronic health records, IEEE Transactions on Neural Networks and Learning Systems 35 (9) (2024) 11843–11856.
21. A. M. Aftab, Representation learning on electronic health records using graph neural networks. N. Ahmed, R. Heyer, D. Broneske, et al., Graph representation of pa-tient's data in ehr for outcome prediction, Ph.D. thesis, Master Thesis, Otto-von-Guericke University (2023).
22. Z. Icten, M. Friedman, J. Menzin, P24 predictors of major adverse car-diovascular events among type 2 diabetes mellitus patients: a machine learning time-to-event analysis, Value in Health 25 (7) (2022) S291–S292.
23. S. Mishra, A comparative study for time-to-event analysis and survival prediction for heart failure condition using machine learning techniques, Journal of Electronics, Electromedical Engineering, and Medical Infor-matics 4 (3) (2022) 115–134.
24. M. Arjovsky, L. Bottou, I. Gulrajani, D. Lopez-Paz, Invariant risk min-imization, arXiv preprint arXiv:1907.02893 (2019).
25. D. Rothenhäusler, N. Meinshausen, P. Bühlmann, J. Peters, Anchor regression: Heterogeneous data meet causality, Journal of the Royal Statistical Society Series B: Statistical Methodology 83 (2) (2021) 215– 246.
26. M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, J. Bian, Causal inference and counter-factual prediction in machine learning for actionable healthcare, Nature Machine Intelligence 2 (7) (2020) 369–375.
27. R. Xu, Y. Sun, C. Chen, P. Venkitasubramaniam, S. Xie, Robust confor-mal prediction under distribution shift via physics-informed structural causal model, arXiv preprint arXiv:2403.15025 (2024).
28. V. Vovk, A. Gammerman, G. Shafer, Algorithmic learning in a random world, Springer Cham, 2005.
29. J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, Journal of the American Statistical Association 113 (523) (2018) 1094–1111.
30. A. N. Angelopoulos, S. Bates, A gentle introduction to conformal pre-diction and distribution-free uncertainty quantification, arXiv preprint arXiv:2107.07511 (2021).

31. N. Dewolf, A comparative study of conformal prediction methods for valid uncertainty quantification in machine learning, arXiv preprint arXiv:2405.02082 (2024).
32. Y. Jin, E. J. Candès, Selection by prediction with conformal p-values, Journal of Machine Learning Research 24 (244) (2023) 1–41.
33. L. Lei, E. J. Candès, Conformal inference of counterfactuals and indi-vidual treatment effects, Journal of the Royal Statistical Society Series B: Statistical Methodology 83 (5) (2021) 911–938.
34. M. Schröder, D. Frauen, J. Schweisthal, K. Heß, V. Melnychuk, S. Feuer-riegel, Conformal prediction for causal effects of continuous treatments, arXiv preprint arXiv:2407.03094 (2024).
35. S. Lolak, J. Attia, G. J. McKay, A. Thakkinstian, Application of drag-onnet and conformal inference for estimating individualized treatment effects for personalized stroke prevention: Retrospective cohort study, JMIR cardio 9 (2025) e50627.
36. S. Liang, Y. Xiao, L. Kong, W. Tang, Adaptive conformal prediction in-tervals for invariant learning, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, 2025, pp. 1695–1706.
37. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282.
38. S. Ghosh, S. K. Ghosh, Feel: Federated learning framework for elderly healthcare using edge-iomt, IEEE Transactions on Computational Social Systems 10 (4) (2023) 1800–1809.
39. S. R. Abbas, Z. Abbas, A. Zahir, S. W. Lee, Federated learning in smart healthcare: a comprehensive review on privacy, security, and predictive analytics with iot integration, in: Healthcare, Vol. 12, MDPI, 2024, p. 2587.
40. X. Wang, J. Hu, H. Lin, W. Liu, H. Moon, M. J. Piran, Federated learning-empowered disease diagnosis mechanism in the internet of med-ical things: From the privacy-preservation perspective, IEEE Transac-tions on Industrial Informatics 19 (7) (2022) 7905–7913.
41. D. R. Birari, K. D. Bamane, P. B. Kamble, T. A. Dhaigude, R. B. Shendge, A. Dandavate, Towards a holistic approach to chronic disease management: Integrating federated learning and iot for personalized health care., Journal of Electrical Systems 19 (3) (2023).
42. D. Patil, S. Goel, S. K. Garud, M. D. Kokate, A. Nashte, P. Rane, Federated learning in real-time medical iot: Optimizing privacy and accuracy for chronic disease monitoring., Journal of Electrical Systems 19 (3) (2023).
43. C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sen-sitivity in private data analysis, in: Theory of cryptography conference, Springer, 2006, pp. 265–284.
44. O. Choudhury, A. Gkoulalas-Divanis, T. Salonidis, I. Sylla, Y. Park, G. Hsu, A. Das, Differential privacy-enabled federated learning for sen-sitive health data, arXiv preprint arXiv:1910.02578 (2019).
45. T. U. Islam, R. Ghasemi, N. Mohammed, Privacy-preserving federated learning model for healthcare data, in: 2022 IEEE 12th Annual Com-puting and Communication Workshop and Conference (CCWC), IEEE, 2022, pp. 0281–0287.
46. B. Li, H. Gao, X. Shi, Feddp: secure federated learning with differential privacy for disease prediction, in: International Conference on Com-putational Advances in Bio and Medical Sciences, Springer, 2023, pp. 119–131.
47. W. Moulahi, T. Moulahi, I. Jdey, S. Zidi, Federated learning harnessed with differential privacy for heart disease prediction: Enhancing privacy and accuracy.
48. S. M. Bokhari, S. Sohaib, M. Shafi, Fusion of personalized federated learning (pfl) with differential privacy (dp) learning for diagnosis of ar-rhythmia disease, Plos one 20 (7) (2025) e0327108.
49. G. Ayyappan, V. Loganathan, E. Padma, S. Ilavarasan, A. Subash, et al., Federated learning and edge ai for privacy-preserving diabetes prediction in healthcare, in: 2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), IEEE, 2025, pp. 1116–1121.
50. G. M. Nagamani, C. K. Kumar, Design of an improved graph-based model for real-time anomaly detection in healthcare using hybrid cnn-lstm and federated learning, Heliyon 10 (24) (2024).
51. L. Itu, S. Rapaka, T. Passerini, B. Georgescu, C. Schwemmer, M. Schoe-binger, T. Flohr, P. Sharma, D. Comaniciu, A machine-learning ap-proach for computation of fractional flow reserve from coronary com-puted tomography, Journal of applied physiology 121 (1) (2016) 42–52.
52. C.-A. Hatfaludi, I.-A. Tache, C. F. Cius, del, A. Puiu, D. Stoian, L. M. Itu, L. Calmac, N.-M. Popa-Fotea, V. Bataila, A. Scafa-Udriste, To-wards a deep-learning approach for prediction of fractional flow reserve from optical coherence tomography, Applied Sciences 12 (14) (2022) 6964.
53. M. Vardhan, A. Randles, Application of physics-based flow models in cardiovascular medicine: Current practices and challenges, Biophysics Reviews 2 (1) (2021) 011302.
54. F. Zuo, C. Dai, L. Wei, Y. Gong, C. Yin, Y. Li, Real-time amplitude spectrum area estimation during chest compression from the ecg wave-form using a 1d convolutional neural network, Frontiers in Physiology 14 (2023) 1113524.
55. C. Wanner, M. Tonelli, et al., Kdigo clinical practice guideline for lipid management in ckd: summary of recommendation statements and clin-ical approach to the patient, Kidney international 85 (6) (2014) 1303– 1309.