SURVBRIDGE: SURVIVAL REPRESENTATION TRANSFER ACROSS HETEROGENEOUS CLINICAL COHORTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterogeneous clinical cohorts often differ in available variables, observation processes, and follow-up, complicating the transfer of survival representations learned from large electronic health records (EHRs). We introduce SurvBridge, a schema-flexible framework for cross-cohort survival representation transfer that combines feature-semantic encoding, survival-aware multi-source contrastive pre-training, explicit representation of missingness state, and Cox-based survival adaptation. SurvBridge supports pre-training across multiple EHR sources with and without time-to-event labels, without requiring identical feature spaces, and adapts the learned representation to target-native clinical schemas. We evaluate SurvBridge on three public clinical survival datasets and two independent cardiovascular cohorts. SurvBridge achieves strong survival discrimination against classical, deep survival, and tabular foundation-model baselines. Pre-training provides its largest gains when labeled target data are scarce, with SurvVPCL providing more consistent low-resource transfer than the alternative survival-aware contrastive objectives evaluated. On naturally incomplete clinical data, explicitly representing missingness state improves mean discrimination over deterministic missing-value handling and changes how partially observed clinical variables contribute to predicted risk. These findings position cross-cohort survival pre-training primarily as a means of improving sample efficiency while accommodating heterogeneous clinical feature spaces and missingness representation.

1 INTRODUCTION

Electronic health records (EHRs) provide large-scale patient data that can support representation learning for clinical prediction. In contrast, developing survival models for specialized clinical cohorts is often limited by the availability of sufficiently large outcome-labeled datasets. This creates an opportunity for transfer learning: representations learned from large external EHR resources could reduce the amount of labeled target data required for survival prediction. However, source EHRs and target cohorts often arise from different clinical and data-generation processes (Goldstein et al., 2016). They may contain different variables, follow patients over different time horizons, and exhibit different observation and missingness patterns (Wan et al., 2022; Subbaswamy & Saria, 2020). Consequently, effective transfer must therefore accommodate these cohort differences rather than assume a common input space or survival process.

Three issues are particularly challenging for cross-cohort survival transfer. First, source and target schemas may overlap only partially, making a fixed feature space unnecessarily restrictive (Wang & Sun, 2022a). Second, differences in clinical setting, outcome definition, and follow-up can make absolute survival times difficult to compare across cohorts, complicating the use of source survival supervision during pre-training. Third, missingness in routinely collected clinical data can itself reflect the observation process, so replacing all unobserved measurements with deterministic imputations (e.g., median filling, constant filling) may discard useful information (Groenwold, 2020; Tan et al., 2023). These challenges are especially consequential when labeled target survival data are scarce, where effective transfer could provide the greatest benefit.

This motivates the central question of our study: *Can heterogeneous external EHR data be used to reduce dependence on labeled target data for clinical survival prediction?* We investigate three research questions. **RQ1:** Can survival representations transfer across heterogeneous clinical cohorts,

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.

and does multi-source pre-training improve downstream survival prediction? **RQ2:** How does the benefit of pre-training change as the amount of labeled target survival data increases? **RQ3:** Does explicitly representing missingness improve survival prediction under naturally occurring clinical missingness?

To address these questions, we introduce **SurvBridge**, a schema-flexible and missingness-aware framework for cross-cohort survival representation transfer. Building on TransTab-style feature-semantic encoding (Wang & Sun, 2022a), SurvBridge maps heterogeneous clinical tables into a shared latent space without requiring identical source and target schemas. It introduces survival-aware multi-source contrastive pre-training across cohorts with and without time-to-event labels, explicitly represents missing observations with a learned state, and adapts the resulting representation to right-censored target outcomes using a Cox proportional-hazards objective.

Our main contributions are:

1. **Survival representation transfer across heterogeneous clinical cohorts.** We develop SurvBridge, a schema-flexible framework for transferring representations across clinical cohorts, enabling adaptation to target clinical schemas without requiring identical source and target variables.
2. **Survival-aware multi-source pre-training.** We introduce SurvVPCL, an extension of vertical-partition contrastive learning that incorporates source survival information through relative event-time proximity while preserving VPCL's same-patient cross-view structure. Across two independent cardiovascular targets, pre-training provides its largest gains when labeled target data are scarce, with VPCL and SurvVPCL providing the most consistent low-resource transfer among the evaluated pre-training objectives.
3. **Missingness-aware clinical transfer.** We explicitly represent missing observations with a learned state and evaluate its transfer across cohorts with heterogeneous observation patterns. The learned representation consistently improves mean discrimination over deterministic imputation strategies and changes the contribution of partially observed clinical variables to predicted risk.

2 Related Work

Survival analysis and deep survival modeling. Classical survival models account for right censoring through approaches such as the Cox proportional-hazards model (Cox, 1972) and Random Survival Forests (Ishwaran et al., 2008). DeepSurv (Katzman et al., 2018) extends Cox modeling with a nonlinear neural risk function, while DeepHit (Lee et al., 2018) directly models discrete-time event distributions. More recent approaches use latent-variable and Transformer-based architectures for survival prediction (Mesinovic et al., 2026; Wang & Sun, 2022b). These methods provide strong survival learners but generally operate within a fixed feature space and do not directly address representation transfer between cohorts with partially overlapping clinical schemas.

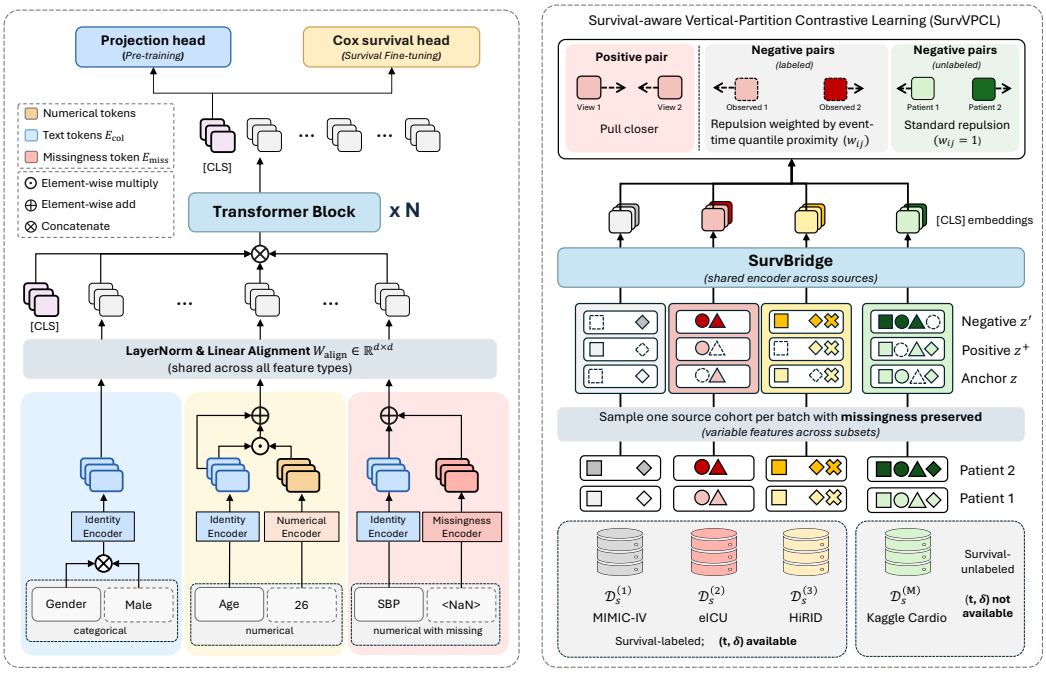
Survival representation learning and transfer. Representation-learning methods such as TransTab (Wang & Sun, 2022a), TabHealth (Campana et al., 2026), and TransEHR (Zhao et al., 2026) support transfer across heterogeneous tabular, longitudinal temporality or clinical feature spaces, but do not specifically address survival-aware multi-source pre-training. Complementarily, survival-aware contrastive methods such as OTCSurv (Nayebi Kerdabadi et al., 2023) and SurvRNC (Saeed et al., 2024) use time-to-event information to structure learned representations, but were introduced for supervised survival learning rather than transfer across heterogeneous cohorts. We study the intersection of these settings: survival-aware representation pre-training across cohorts that may differ in feature spaces, follow-up scales, and availability of survival labels. Recent tabular foundation models provide an alternative approach to limited-data survival prediction (Pham et al., 2026b;a; ang Qi et al., 2026); we include these as downstream benchmarks while focusing on transferable clinical representation learning.

Cross-cohort heterogeneity and clinical missingness. Differences in patient populations, clinical practice, follow-up, and observation processes can limit model transferability even when variables are semantically harmonized (Wiens et al., 2014). Recent multi-cohort evaluation further demonstrates variation in model behavior across clinical datasets with different scales, feature spaces, and

outcomes (Pham et al., 2026c). Missingness introduces an additional source of heterogeneity because whether a measurement is observed can itself carry predictive information (Groenwold, 2020), while observation patterns may differ substantially across cohorts. We therefore study whether explicitly representing observation state during pre-training transfers more effectively than deterministic imputation under cross-cohort missingness shift.

3 METHODOLOGY

We propose SurvBridge, a schema-flexible survival representation framework for transferring representations across heterogeneous clinical tables. The model combines feature-semantic encoding, explicit representation of patient-level missingness, survival-aware multi-source contrastive pre-training, and survival-specific target adaptation. Figure 1 summarizes the framework, with Figure 1a illustrating the feature encoding and target adaptation architecture and Figure 1b illustrating the multi-source pre-training procedure.



(a) SurvBridge architecture.

(b) Multi-source survival pre-training.

Figure 1: SurvBridge framework for cross-cohort survival representation transfer. (a) Schema-flexible feature encoding with an explicit learned representation of missing observations, followed by target survival adaptation. (b) Multi-source contrastive pre-training with a shared encoder. SurvVPCL preserves same-patient cross-view positives while using relative event-time proximity to modulate the repulsion between patients from survival-labeled sources; sources without survival labels reduce to standard VPCL.

3.1 PROBLEM FORMULATION & SURVIVAL NOTATION

We consider right-censored survival data of N patients: $\mathcal{D} = \{(\mathbf{x}_i, \tilde{T}_i, \delta_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ denotes the covariates of patient i, $T_i \geq 0$ is the event time, $C_i \geq 0$ is the censoring time, $\tilde{T}_i = \min(T_i, C_i)$ is the observed follow-up time, and $\delta_i = \mathbb{I}(T_i \leq C_i)$ indicates whether the event was observed. The conditional survival function is

$$S(t \mid \mathbf{x}) = \Pr(T > t \mid \mathbf{x}), \qquad t \geq 0. \tag{1}$$

For downstream prediction, we adopt the Cox proportional hazards formulation

$$h(t \mid \mathbf{x}_i) = h_0(t) \exp(g_\theta(\mathbf{x}_i)), \qquad t \geq 0, \tag{2}$$

where $h_0(t) \geq 0$ is the unspecified baseline hazard and $g_\theta(\mathbf{x}_i) \in \mathbb{R}$ is the subject-specific log-relative hazard.

3.2 Schema-Flexible Clinical Feature Encoding

Following TransTab (Wang & Sun, 2022a), SurvBridge represents clinical variables by their feature identity rather than fixed column position, allowing a shared encoder to process partially overlapping feature spaces. Feature identities are represented through their textual descriptions and encoded using BioLORD (Remy et al., 2022). Dataset-specific abbreviations or opaque feature names can optionally be expanded to more descriptive clinical terms (e.g., `SBP` to `systolic blood pressure`) before encoding, but no one-to-one mapping between source and target features is required. BioLORD token embeddings are mean-pooled and mapped to the Transformer hidden dimension ($d = 128$) through a shared learnable projection, yielding the feature-identity representation $E_{\text{col}}(c)$. The BioLORD encoder is kept frozen, while the shared projection is learned jointly with the remaining model parameters.

3.3 Missingness-Aware Scalar Embeddings

As illustrated in Figure 1a, SurvBridge explicitly preserves whether each scalar measurement is observed or missing. For each feature c, we define the observation state for patient i as $m_{i,c} = \mathbb{I}\left(x_{i,c} \text{ is missing}\right)$. Within SurvBridge, missing scalar values are not imputed. Instead, SurvBridge represents an unobserved measurement with a learned missingness embedding. Observed scalar values are standardized and incorporated through a projection, whereas missing values bypass this numerical projection entirely.

The resulting numerical feature token is

$$E_{\text{num}}(i,c) = \begin{cases} E_{\text{col}}(c) + \alpha\left(E_{\text{col}}(c)\tilde{x}_{i,c} + B\right), & m_{i,c} = 0, \\ E_{\text{col}}(c) + E_{\text{miss}}, & m_{i,c} = 1, \end{cases} \tag{3}$$

where $\tilde{x}_{i,c}$ denotes the standardized observed value, α is a learned scalar, B is a learned numerical bias, and E_{miss} is a learned embedding vector representing the missing observation state.

Categorical variables are represented by concatenating the expanded feature name with the observed category value before biomedical text encoding, yielding $E_{\text{cat}}(i,c)$. If the categorical value is missing, the feature name is retained while the category value is represented by an empty string.

Before entering the Transformer, each numerical or categorical feature token $E(i,c) \in \{E_{\text{num}}(i,c), E_{\text{cat}}(i,c)\}$ is normalized and passed through a shared linear alignment layer,

$$\widetilde{E}(i,c) = W_{\text{align}}\ \text{LN}(E(i,c)), \qquad W_{\text{align}} \in \mathbb{R}^{d\times d}, \tag{4}$$

where $d = 128$ and the alignment parameters are shared across feature types and features.

The aligned feature tokens are then concatenated along the token dimension, and a learned classification token, [CLS], is prepended to the sequence. The resulting patient-level input is

$$\mathbf{E}_i = E_{\text{CLS}} \oplus \left\{\widetilde{E}_{\text{num}}(i,c)\right\}_{c\in\mathcal{F}^{\text{num}}} \oplus \left\{\widetilde{E}_{\text{cat}}(i,c)\right\}_{c\in\mathcal{F}^{\text{cat}}}.$$

The final hidden state corresponding to [CLS] is used as the patient representation for contrastive pre-training and downstream survival prediction.

3.4 Semi-Supervised Multi-source Survival Contrastive Pre-training

Figure 1b illustrates the multi-source pre-training procedure. SurvBridge builds on vertical-partition contrastive learning (VPCL) to pre-train the shared encoder before survival adaptation. For each patient, the available features are partitioned into K overlapping subsets with overlap ratio p. Each view is independently encoded by the shared Transformer, and its [CLS] representation is mapped through a projection head to $z_i^{(v)} \in \mathbb{R}^{d_{\text{proj}}}$. Views of the same patient form positive pairs (z_i, z_i^+),

while views from different patients form negatives (z_i, z_j). We write the contrastive objective as

$$\mathcal{L}_{\text{SurvVPCL}} = -\mathbb{E}_{(z_i, z_i^+)} \left[\log \frac{\exp\big(\text{sim}(z_i, z_i^+)/\tau\big)}{\exp\big(\text{sim}(z_i, z_i^+)/\tau\big) + \sum_{j \neq i} w_{ij} \exp(\text{sim}(z_i, z_j)/\tau)} \right], \quad (5)$$

where $\text{sim}(z_i, z_j) = z_i^\top z_j$, τ is the temperature, and w_{ij} controls the contribution of each inter-patient negative. Standard VPCL corresponds to $w_{ij} = 1$ for all $i \neq j$.

Survival-Weighted Negatives. Inspired by OTCSurv (Nayebi Kerdabadi et al., 2023), we use temporal proximity to incorporate survival information during contrastive pre-training; unlike OTCSurv, however, survival-aware VPCL (SurvVPCL) preserves VPCL's same-patient cross-view positives and uses survival information only to reweight inter-patient negatives. For an observed-event patient i from source s, event time t_i is mapped to its within-source event-time quantile $u_i = \widehat{F}_{s,\text{event}}(t_i) \in [0, 1]$, and we define

$$w_{ij} = 1 - \mathbb{I}(\delta_i = 1, \delta_j = 1) \exp\left(-\frac{|u_i - u_j|}{\sigma} \right), \quad (6)$$

where δ_i indicates an observed event and σ controls the temporal neighborhood. Thus, patients with nearby observed event-time quantiles remain negative pairs but contribute less strongly to the contrastive denominator (i.e., they are more weakly repelled; Figure 1b). As temporal distance increases, w_{ij} approaches one and the pair approaches standard VPCL weighting. Pairs involving a censored patient retain $w_{ij} = 1$. Using within-source event-time quantiles avoids directly comparing absolute survival times across cohorts with different follow-up scales.

Semi-Supervised Multi-Source Pre-training. Mini-batches are drawn from one source at a time while encoder parameters are shared across sources. Sources with time-to-event labels use the survival-weighted objective in Eq. 5, whereas unlabeled sources set $w_{ij} = 1$ and therefore reduce exactly to standard VPCL. This allows both survival-labeled and survival-unlabeled cohorts to contribute jointly to the same shared representation.

3.5 Survival Adaptation

For target survival prediction, the contrastive projection head is replaced by a survival-specific risk network comprising layer normalization, a two-layer MLP with ReLU activations and dropout, and a linear output layer. This network is applied to the final Transformer [CLS] representation and produces the scalar log-risk score $g_\theta(\mathbf{x}_i)$ for patient i. The encoder and survival head are then jointly optimized using the Cox negative partial log-likelihood Cox (1972)

$$\mathcal{L}_{\text{Cox}}(\theta) = -\sum_{i:\delta_i=1} \left[g_\theta(\mathbf{x}_i) - \log \sum_{j \in \mathcal{R}(\tilde{T}_i)} \exp\left(g_\theta(\mathbf{x}_j)\right) \right], \quad (7)$$

where $\mathcal{R}(\tilde{T}_i) = \left\{ j : \tilde{T}_j \geq \tilde{T}_i \right\}$ is the risk set at the observed event time of patient i.

After fine-tuning, we estimate the target-specific baseline survival function $\hat{S}_0(t)$ from the corresponding training partition and obtain individual survival estimates as

$$\hat{S}(t \mid \mathbf{x}_i) = \hat{S}_0(t)^{\exp(g_\theta(\mathbf{x}_i))}. \quad (8)$$

4 Experimental Design

4.1 Clinical Cohorts and Outcomes

We evaluate SurvBridge across heterogeneous clinical cohorts that differ in population, feature availability, observation process, and follow-up. For multi-source pre-training, we use MIMIC-IV (Johnson et al., 2023), eICU (Pollard et al., 2018), HiRID (Yèche et al., 2021), and KaggleCD, an external

cardiovascular disease dataset[1], which differ substantially in feature availability, clinical setting, and observation process.

We evaluate downstream performance on two independent private cardiovascular cohorts, URRAH and IHD , both providing right-censored time-to-event outcomes (source details omitted for double-blind review, they will be added to camera ready version). Additionally, we benchmark our methods on 3 public clinical datasets derived from SurvSet Drysdale (2022) (SUPPORT2, SMARTO, FLCHAIN). Complete cohort characteristics and feature statistics are provided in Appendix Table 2.

4.2 Evaluation Protocols

We compare SurvBridge against classical survival models (CoxPH, RSF, and GBSA), deep survival models (DeepSurv, DeepHit, DSM, DySurv, and SurvTRACE), TransTab with a Cox head, and tabular foundation-model baselines (SurvivalPFN, TabPFN-Cox, and TabPFN-MTLR). All baselines use only target-cohort training data with median imputation. SurvBridge is evaluated both from scratch and after multi-source pre-training on MIMIC-IV, eICU, HiRID, and KaggleCD. For pre-training, we compare VPCL and the proposed SurvVPCL with the survival-aware objectives OTC-Surv (Nayebi Kerdabadi et al., 2023) and SurvRNC (Saeed et al., 2024); implementation details are provided in the Appendix.

For low-resource transfer, we use target training subsets of $N \in \{100, 200, 500, 1000, 2000, 4000, 6000, 10000\}$ patients, subject to cohort size, with matched subsets and fixed validation and test partitions across methods. For missingness analysis, we compare median and mean imputation, zero filling, and the learned missingness representation on naturally incomplete URRAH data, using matched handling during pre-training and target adaptation. We also use SurvSHAP (Krzyziński et al., 2023) to examine changes in feature importance rankings across approaches.

We use a fixed held-out test partition and five random seeds with five train–validation folds. Fold-specific predictions are aggregated per seed, and we report mean C^{td} (Antolini et al., 2005) with 95% confidence intervals across seeds. Matched seeds and subsets are used for paired comparisons. Non-PFN baselines use 30 hyperparameter trials. Complete preprocessing, optimization, evaluation, and search-space details are provided in the Appendix, together with various experiments motivating methodological choices and additional results compared to the main manuscript.

5 Results

We evaluate SurvBridge along three complementary axes: overall survival performance, transfer under limited target supervision, and missingness-aware representation.

Cross-Cohort Performance. Across the five target cohorts, SurvBridge achieves strong survival discrimination relative to the evaluated classical, deep survival, and foundation-model baselines (Table 1). Importantly, this performance is not solely attributable to source pre-training. When trained from scratch, SurvBridge already achieves competitive or higher mean C^{td} than the evaluated baselines on the two cardiovascular targets (URRAH and IHD), demonstrating the benefit of the survival-specific, schema-flexible architecture independently of representation transfer.

Pre-training provides an additional benefit across the five cohorts. On the three public SurvSet-derived targets, SurvVPCL achieves the highest mean performance among the SurvBridge variants, reaching 0.778, 0.671, and 0.795 on SUPPORT2, SMARTO, and FLCHAIN, respectively, compared with 0.776, 0.665, and 0.794 for standard VPCL. Although these differences are modest, their consistent direction across the three public cohorts indicates that survival-aware weighting can complement VPCL while preserving its cross-cohort transfer behavior. On URRAH and IHD, VPCL and SurvVPCL perform similarly under full target supervision, motivating the low-resource analysis below.

Pre-training Improves Sample Efficiency. Figure 2 shows that representation pre-training provides its largest benefit when target supervision is limited. At $N = 100$, SurvVPCL improves

[1] https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data

Table 1: Main predictive performance across target and public survival cohorts, measured by time-dependent concordance index (C^{td} ↑). Mean and 95% confidence intervals are computed across up to 5 seeds; the interval is shown as a subscript. Best and second-best means within each dataset are **bolded** and underlined, respectively.

Model	URRAH	IHD	SUPPORT2	SMARTO	FLCHAIN
Classical survival models					
CoxPH	$0.822_{[0.815, 0.828]}$	$0.797_{[0.787, 0.807]}$	$0.769_{[0.758, 0.780]}$	$0.650_{[0.612, 0.679]}$	$0.790_{[0.777, 0.804]}$
RSF	$0.829_{[0.817, 0.840]}$	$0.799_{[0.790, 0.807]}$	$0.766_{[0.751, 0.781]}$	$0.637_{[0.603, 0.672]}$	$0.775_{[0.760, 0.789]}$
GBSA	$0.831_{[0.824, 0.838]}$	$0.795_{[0.784, 0.805]}$	$0.753_{[0.747, 0.760]}$	$0.656_{[0.634, 0.677]}$	$0.792_{[0.779, 0.805]}$
Deep survival models					
DeepSurv	$0.818_{[0.765, 0.870]}$	$0.798_{[0.787, 0.809]}$	$0.770_{[0.762, 0.777]}$	$0.654_{[0.605, 0.703]}$	$0.792_{[0.779, 0.805]}$
DeepHit	$0.838_{[0.831, 0.845]}$	$0.780_{[0.772, 0.789]}$	$0.752_{[0.745, 0.760]}$	$0.556_{[0.527, 0.586]}$	$0.759_{[0.751, 0.767]}$
DySurv	$0.824_{[0.808, 0.841]}$	$0.787_{[0.774, 0.799]}$	$0.741_{[0.728, 0.753]}$	$0.649_{[0.618, 0.681]}$	$0.790_{[0.775, 0.805]}$
SurvTRACE	$0.851_{[0.843, 0.858]}$	$0.760_{[0.756, 0.763]}$	$0.702_{[0.689, 0.714]}$	$0.610_{[0.586, 0.634]}$	$0.780_{[0.765, 0.795]}$
Deep Survival Machine	$0.839_{[0.833, 0.846]}$	$0.799_{[0.792, 0.806]}$	$0.771_{[0.758, 0.783]}$	$0.647_{[0.597, 0.697]}$	$0.792_{[0.777, 0.807]}$
TransTab-Cox	$0.841_{[0.834, 0.848]}$	$0.799_{[0.791, 0.808]}$	$0.703_{[0.682, 0.724]}$	$0.616_{[0.572, 0.661]}$	$0.791_{[0.774, 0.808]}$
In-context tabular foundation models					
SurvivalPFN	$0.739_{[0.724, 0.754]}$	$0.606_{[0.581, 0.631]}$	$0.613_{[0.515, 0.711]}$	$0.513_{[0.491, 0.536]}$	$0.629_{[0.600, 0.659]}$
TabPFN-Cox	$0.840_{[0.834, 0.846]}$	$0.798_{[0.789, 0.807]}$	$0.772_{[0.765, 0.779]}$	$\underline{0.667}_{[0.625, 0.708]}$	$0.793_{[0.780, 0.807]}$
TabPFN-MTLR	$0.848_{[0.842, 0.853]}$	$0.791_{[0.786, 0.796]}$	$0.774_{[0.761, 0.786]}$	$0.642_{[0.596, 0.687]}$	$0.787_{[0.773, 0.802]}$
SurvBridge (Ours)					
SurvBridge (*Scratch*)	$0.855_{[0.849, 0.861]}$	$0.800_{[0.791, 0.810]}$	$0.755_{[0.749, 0.761]}$	$0.651_{[0.613, 0.689]}$	$0.773_{[0.758, 0.788]}$
SurvBridge (*VPCL*)	$\underline{0.858}_{[0.851, 0.865]}$	$\mathbf{0.803}_{[0.790, 0.816]}$	$\underline{0.776}_{[0.768, 0.785]}$	$0.665_{[0.628, 0.702]}$	$\underline{0.794}_{[0.778, 0.810]}$
SurvBridge (*SurvVPCL*)	$\mathbf{0.859}_{[0.853, 0.865]}$	$\underline{0.802}_{[0.788, 0.816]}$	$\mathbf{0.778}_{[0.767, 0.788]}$	$\mathbf{0.671}_{[0.635, 0.707]}$	$\mathbf{0.795}_{[0.780, 0.810]}$

C^{td} over training from scratch by 0.118 on URRAH and 0.088 on IHD, with substantial gains also observed at $N = 200$ and $N = 500$. The advantage progressively narrows as additional target supervision becomes available. VPCL and SurvVPCL exhibit similar transfer performance across sample sizes, with neither uniformly dominating the other, and generally provide stronger and more persistent low-resource transfer than SurvRNC and OTCSurv. These results indicate that the principal benefit of source pre-training is improved sample efficiency, while SurvVPCL incorporates survival information without sacrificing the transferable cross-view structure of VPCL.

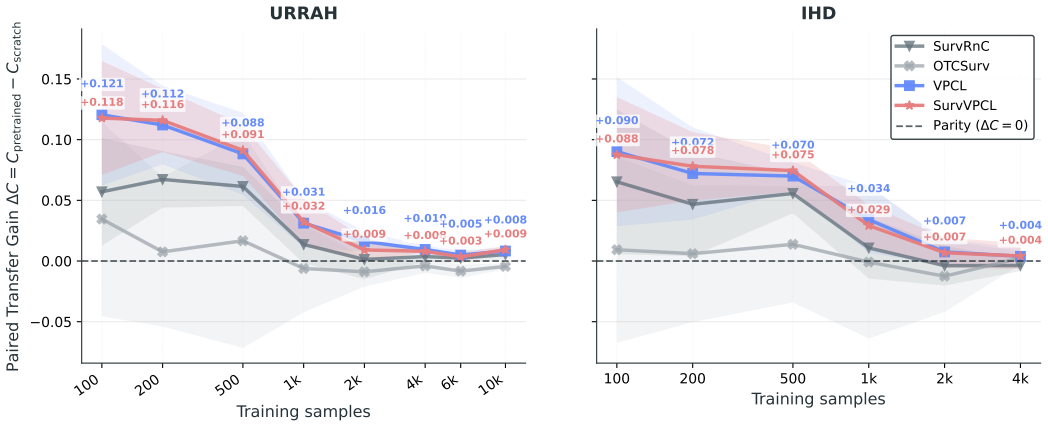


Figure 2: Low-resource transfer on URRAH and IHD. Models are adapted using nested target-training subsets while the held-out test partition remains fixed. Paired gains report the difference in C^{td} between each pre-trained model and the same architecture trained from scratch on the matched target subset.

Missingness-Aware Representation. Figure 3 compares deterministic missing-value handling with the proposed learned missingness representation on naturally incomplete URRAH data. The learned missingness variant achieves the highest mean C^{td} across both full-data and $N = 1000$

settings, with and without matched pre-training. Its difference from zero-constant filling is small in several settings, whereas the advantage over median imputation is more pronounced, particularly in the low-resource setting. Pre-training additionally improves mean performance across the evaluated missing-value handling strategies, with the largest gains occurring at $N = 1000$.

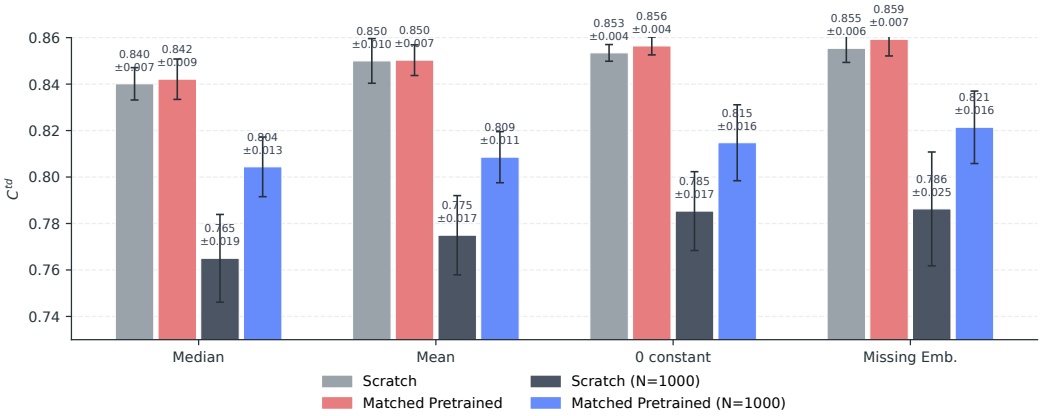


Figure 3: Missingness ablation on naturally incomplete URRAH data, comparing deterministic missing-value handling with the learned missingness representation. Results are shown for models trained from scratch and with matched pre-training under the full-data and $N = 1000$ target-data regimes. Bars show mean C^{td} with 95% Student-t confidence intervals across five seeds.

We further examine how the missingness representation affects model behavior using SurvSHAP. As shown in Figure 4, several highly incomplete variables, including atrial fibrillation, CRP, and allopurinol use, receive greater relative importance under the learned missingness representation than under deterministic imputation. Atrial fibrillation provides an illustrative example: as its importance increases, the importance of the more frequently observed beta-blocker and calcium-channel-blocker (CCB) use decreases. Since these treatments are clinically related to atrial fibrillation, this shift suggests that explicitly representing observation state changes how the model distributes predictive information between the partially observed condition and related treatment variables.

6 Discussion

Our results separate the contributions of the SurvBridge design from those of cross-cohort pre-training. On the two cardiovascular cohorts, SurvBridge trained from scratch is already competitive with strong classical, deep survival, and foundation-model baselines, indicating that its survival-specific and schema-flexible design provides value independently of representation transfer. Pre-training provides further benefit across heterogeneous target cohorts, with SurvVPCL achieving the highest mean performance among the SurvBridge variants on all three public survival benchmarks. Its benefit is most pronounced under limited target supervision: on URRAH and IHD, gains are substantial with only a few hundred labeled patients and progressively narrow as the target training set grows. Cross-cohort pre-training therefore appears particularly valuable for improving sample efficiency when target survival labels are limited.

The comparison of pre-training objectives further shows that survival supervision does not automatically yield large improvements in transfer. In the low-resource experiments on URRAH and IHD, VPCL and SurvVPCL perform similarly and generally provide stronger and more persistent gains than SurvRNC and OTCSurv. Across the three public SurvSet-derived targets, SurvVPCL consistently achieves higher mean performance than VPCL, although the differences are modest. Together, these results suggest that SurvVPCL can incorporate survival information while preserving the transferable cross-view structure of VPCL.

Missingness represents another important source of clinical heterogeneity. On naturally incomplete URRAH data, explicitly representing observation state yields consistently higher mean discrimination than deterministic missing-value handling. SurvSHAP further shows that this representation

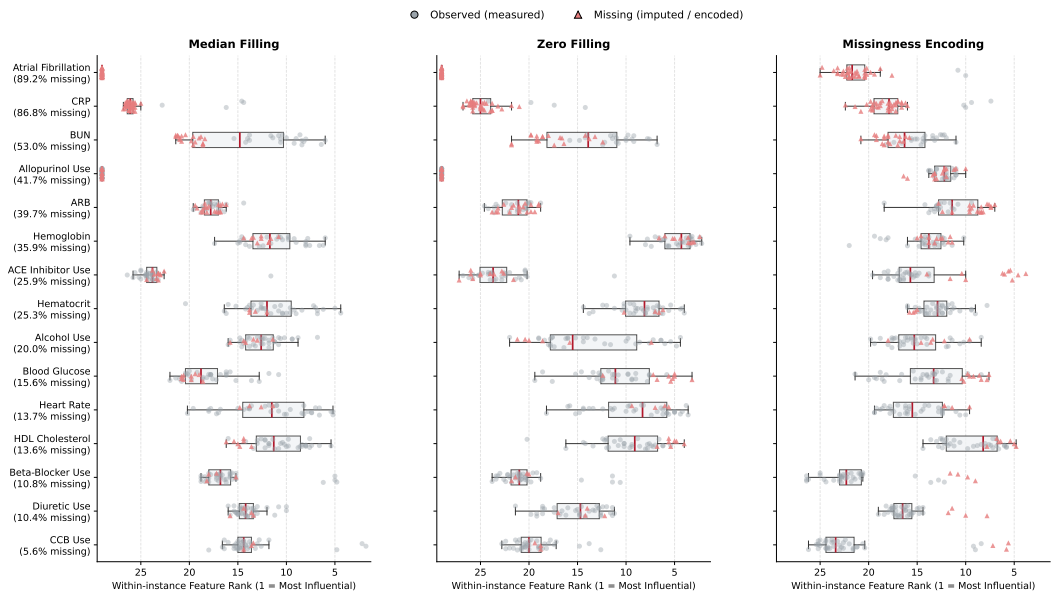


Figure 4: SurvSHAP feature-importance ranks under median filling, zero filling, and learned missingness encoding on URRAH. Features are ordered by their missingness ratio, with the percentage of missing observations shown in parentheses. Lower ranks indicate greater within-instance influence (rank 1 = most influential). Circles denote instances in which the feature is observed, while triangles denote instances in which it is missing and therefore imputed or explicitly encoded. Boxplots summarize the distribution of within-instance ranks.

changes how predictive information is distributed across partially observed and clinically related variables; for example, increased relative importance of atrial fibrillation is accompanied by reduced importance of related treatment variables. These findings support treating observation state as part of the patient representation rather than solely as a preprocessing problem. However, observation patterns may reflect local measurement practices and care pathways as well as patient state, so these attributions should not be interpreted as causal clinical effects.

Several limitations define the scope of these findings. Although the main benchmark includes five target cohorts, detailed low-resource experiments are conducted on URRAH and IHD, and the missingness analysis is limited to URRAH. These findings therefore require validation across additional diseases, endpoints, institutions, and observation processes. Multi-source transfer may also depend on source composition, feature coverage, population similarity, and endpoint compatibility, which are not independently isolated here. Future work should investigate these factors, extend missingness-aware transfer to additional cohorts, and examine source selection or weighting under limited target supervision.

7 Conclusion

We studied survival representation transfer across clinical cohorts with heterogeneous feature spaces, follow-up, observation patterns, and target supervision. Across five target cohorts, SurvBridge achieves strong survival discrimination while operating on target-native feature spaces. Pre-training provides its largest gains when labeled target data are scarce, while explicitly representing observation state improves prediction under natural missingness and changes how partially observed variables contribute to risk. Together, these findings highlight sample efficiency and observation state as central considerations for survival transfer across heterogeneous clinical cohorts.

AI USE STATEMENT

In this work, we used generative AI tools to support research brainstorming, organization of the paper structure, literature discovery, summarization of papers during literature review, code suggestions and debugging assistance, and brainstorming of figures, charts, and diagrams. We did not use generative AI to design the study objectives, produce the core methodological components of SurvBridge, define the survival objectives or evaluation metrics, determine experimental outcomes, or generate the reported results. The manuscript was written and revised by the authors rather than generated by AI.

All AI-assisted material was manually reviewed and verified by the first author. Suggested references and literature summaries were checked against the original publications, and AI-assisted code suggestions were inspected, integrated manually, and validated through the project's testing and experimental pipeline. Figures and analyses were generated from the authors' code and experimental outputs, with AI used only for design or presentation brainstorming where applicable. We take responsibility for the final content of this work, including all text, claims, code, analyses, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study uses the publicly available MIMIC-IV (v3.1) database (Johnson et al., 2023), eICU (v1.0) (Pollard et al., 2018), HiRID (Yèche et al., 2021) accessed via PhysioNet under credentialed agreements.

Among target cohorts, we employ two private de-identified clinical cohorts, IHD and URRAH, accessed through research agreements and processed separately at origin. IHD contains data from IHD patients hospitalized at one of our institutes (to be specified after double-blind review). Data were acquired in the context of institutional clinical assistance within clinical care purposes in a retrospectively collected modality. Informed consent was obtained from each patient. The study complied with the Declaration of Helsinki regarding ethical conduct of research involving human subjects. The URRAH dataset comes from a multicentre retrospective, observational cohort study, which involves data from several cohorts recruited within different centers of hypertension (more details will be provided after double-blind review). Data collection and processing followed the Declaration of Helsinki for Human Research. The processing of the patients' personal data collected complied with the European Directive on the privacy of data, and for this study we obtained access only to anonymous data.

The analysis versions of IHD and URRAH cannot be redistributed under their institutional data-access arrangements.

REPRODUCIBILITY STATEMENT

Section 4 specifies the transfer settings, evaluation protocol, metrics, and target-data regimes. The Appendix provides additional details on the clinical cohorts, preprocessing, model configuration, hyperparameter search spaces, and training procedures. Comparisons use matched data splits and random seeds where applicable. We provide the complete anonymized codebase with the submission, including preprocessing, pre-training, fine-tuning, baseline evaluation, and experiment configurations, together with a reproducibility guide containing environment setup, data preparation, and instructions for reproducing the reported experiments and results. We also provide the trained model parameters for each model type. Access to non-public clinical cohorts remains subject to their respective institutional data-use agreements.

REFERENCES

Shi ang Qi, Vahid Balazadeh, Michael Cooper, Russell Greiner, and Rahul G Krishnan. Survivalpfn: Amortizing survival prediction via in-context bayesian inference. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026. doi: 10.48550/arXiv.2605.15488. URL https://openreview.net/forum?id=PDik7bpFhE.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005. doi: https://doi.org/10.1002/sim.2427.

Mattia G Campana, Giada Anastasi, Stefania Pieroni, Michela Franchini, Sabrina Molinaro, and Franca Delmastro. Tabhealth: Towards a self-supervised tabular foundation model for smart healthcare. In *2026 IEEE International Conference on Smart Computing (SmartComp)*, pp. 24–31. IEEE, 2026. doi: 10.1109/SmartComp69968.2026.00015.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. doi: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Erik Drysdale. Survset: An open-source time-to-event dataset repository. *arXiv preprint arXiv:2203.03094*, 2022. doi: 10.48550/arXiv.2203.03094.

Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John PA Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 24(1):198, 2016. doi: doi:10.1093/jamia/ocw042.

Rolf H. H. Groenwold. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and Prognostic Research*, 4(1):8, 2020. doi: 10.1186/s41512-020-00077-0.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008. doi: http://dx.doi.org/10.1214/08-AOAS169.

Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Beth E Landon, Leo A Celi, and Roger G Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: https://doi.org/10.1038/s41597-022-01899-x.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018. doi: https://doi.org/10.1186/s12874-018-0482-1.

Mateusz Krzyziński, Mikołaj Spytek, Hubert Baniecki, and Przemysław Biecek. Survshap(t): Time-dependent explanations of machine learning survival models. *Knowledge-Based Systems*, 262:110234, 2023. doi: 10.1016/j.knosys.2022.110234.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11842.

Munib Mesinovic, Peter Watkinson, and Tingting Zhu. Dysurv: dynamic deep learning model for survival analysis with conditional variational inference. *Journal of the American Medical Informatics Association*, 33(1):112–122, 2026. doi: 10.1093/jamia/ocae271.

Mohsen Nayebi Kerdabadi, Arya Hadizadeh Moghaddam, Bin Liu, Mei Liu, and Zijun Yao. Contrastive learning of temporal distinctiveness for survival analysis in electronic health records. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 1897–1906, 2023. doi: 10.1145/3583780.3614824.

Minh-Khoi Pham, Luca Cotugno, Dan Cernei, Alina Sirbu, Stefano Masi, Giuseppe Prencipe, Alessandro Pingitore, Patrizia Landi, Tai Tan Mai, Martin Crane, et al. Adaptation interfaces for in-context tabular foundation models in time-to-event prediction. *arXiv preprint arXiv:2609.04901*, 2026a. doi: 10.48550/arXiv.2609.04901.

Minh-Khoi Pham, Luca Cotugno, Alina Sîrbu, Tai Tan Mai, Martin Crane, and Marija Bezbradica. Tabular foundation models for clinical survival analysis via survival-aware adaptation. In *International Conference on AI in Healthcare*, pp. 315–328. Springer, 2026b. doi: https://doi.org/10.1007/978-3-032-35387-0_23.

Minh-Khoi Pham, Thang-Long Nguyen Ho, Thao Thi Phuong Dao, Tai Tan Mai, Minh-Triet Tran, Marie E Ward, Una Geary, Rob Brennan, Nick McDonald, Martin Crane, and Marija Bezbradica. Retrieval-aligned tabular foundation models enable robust clinical risk prediction in electronic health records under real-world constraints. *arXiv preprint arXiv:2604.01841*, 2026c. doi: https://doi.org/10.48550/arXiv.2604.01841.

Tom J Pollard, Alistair E W Johnson, Jesse D Raffa, Leo Anthony Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, 2018. doi: 10.1038/sdata.2018.178.

François Remy, Kris Demuynck, and Thomas Demeester. Biolord: learning ontological representations from definitions for biomedical concepts and their textual descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1454–1465, 2022. doi: 10.18653/v1/2022.findings-emnlp.104.

Numan Saeed, Muhammad Ridzuan, Fadillah Adamsyah Maani, Hussain Alasmawi, Karthik Nandakumar, and Mohammad Yaqub. Survrnc: Learning ordered representations for survival prediction using rank-n-contrast. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15005, pp. 659–669. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72086-4_62.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020. doi: 10.1093/biostatistics/kxz041.

Amelia L. M. Tan, Emily J. Getzen, Meghan R. Hutch, Zachary H. Strasser, Alba Gutiérrez-Sacristán, Trang T. Le, Arianna Dagliati, et al. Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? *Journal of Biomedical Informatics*, 139:104306, 2023. doi: 10.1016/j.jbi.2023.104306.

Bohua Wan, Brian Caffo, and S Swaroop Vedula. A unified framework on generalizability of clinical prediction models. *Frontiers in Artificial Intelligence*, 5:872720, 2022. doi: 10.3389/frai.2022.872720.

Zicheng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:2902–2915, 2022a. doi: https://doi.org/10.52202/068431-0210.

Zifeng Wang and Jimeng Sun. Survtrace: Transformers for survival analysis with competing events. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–9, 2022b. doi: 10.1145/3535508.3545521.

Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4):699–706, 2014. doi: 10.1136/amiajnl-2013-002162.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Ratsch. Hirid-icu-benchmark–a comprehensive machine learning benchmark on high-resolution icu data. In *Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021. doi: 10.48550/arXiv.2111.08536. URL https://openreview.net/forum?id=SnC9rUeqiqd.

Xiongjun Zhao, Zhengyu Liu, Boya Ji, Peng Xi, and Shaoliang Peng. Transehr: Alignment-free electronic health records continual learning across feature spaces. *Expert Systems with Applications*, 296:129020, 2026. doi: 10.1016/j.eswa.2025.129020.

A Appendix

A.1 Clinical Cohorts

Table 2: Summary of datasets used for multi-source pre-training and downstream survival fine-tuning.

Cohort	No. Patients	Event Rate	No. Features	Missingness Rate
Pre-training datasets				
MIMIC-IV	223,380	5.26%	89	28.85%
eICU	139,367	10.64%	88	22.13%
HiRID	33,905	6.08%	24	79.16%
KaggleCD *(Unlabeled)*	68,636	–	11	0.00%
Fine-tuning datasets				
URRAH	27,078	14.84%	63	39.59%
IHD	8,065	29.84%	78	1.08%
SUPPORT2	9,000	68.11%	35	10,9%
SMARTO	3,873	11.88%	26	6.7%
FLCHAIN	7,874	27.52%	10	1.7%

Pre-training cohorts. Table 2 summarizes the characteristics of the pre-training and downstream cohorts used throughout the study. We construct the pre-training pool from four complementary clinical datasets: MIMIC-IV Johnson et al. (2023), eICU Pollard et al. (2018), HiRID Yèche et al. (2021), and KaggleCD. MIMIC-IV and eICU provide heterogeneous inpatient and critical-care data, with eICU spanning a multi-center ICU network. HiRID provides high-resolution critical-care data, while KaggleCD contains ambulatory cardiovascular screening data centered on common cardiovascular risk factors. Together, these sources span distinct patient populations, care settings, feature spaces, and observation processes.

Fine-tuning cohorts. We evaluate downstream transfer on five survival cohorts. IHD and URRAH are two independent cardiovascular cohorts containing detailed clinical, laboratory, and treatment information; URRAH additionally retains naturally occurring patient-level missingness and is therefore used for the missingness-aware representation experiments. We further evaluate on three public clinical survival benchmarks from SurvSet (Drysdale, 2022): SUPPORT2, SMARTO, and FLCHAIN. Together, these cohorts provide heterogeneous target populations, feature spaces, and survival outcomes for evaluating cross-cohort transfer.

A.2 Motivation for the Data Choice and Training Framework of SurvBridge

The clinical cohorts considered in this study differ not only in their available features but also in their empirical covariate and follow-up distributions. Figure 5 shows substantial covariate heterogeneity across the pre-training and target cohorts. Even after harmonizing feature definitions and units, variables including age, blood pressure, hemodynamic measurements, lipid markers, and renal markers differ considerably in location, spread, and distributional shape across datasets. This motivates a schema-flexible shared encoder and multi-source pre-training framework that can learn from heterogeneous feature spaces and empirical distributions rather than assuming a common source and target data distribution.

The cohorts also differ markedly in their follow-up distributions (Figure 6), reflecting differences in clinical setting and outcome definition. Consequently, absolute event times are not directly comparable across pre-training sources. This motivates the within-source event-time quantile transformation used by SurvVPCL, which represents event times on a common relative scale before computing the survival-weighted negative weights in Eq. 6.

Missingness patterns also vary substantially across cohorts, even for features that are shared with the target dataset. Figure 7 orders URRAH features by their missingness rate and compares the corresponding rates whenever the same features are available in the pre-training sources. The resulting patterns are highly cohort dependent: a variable that is frequently observed in URRAH may be

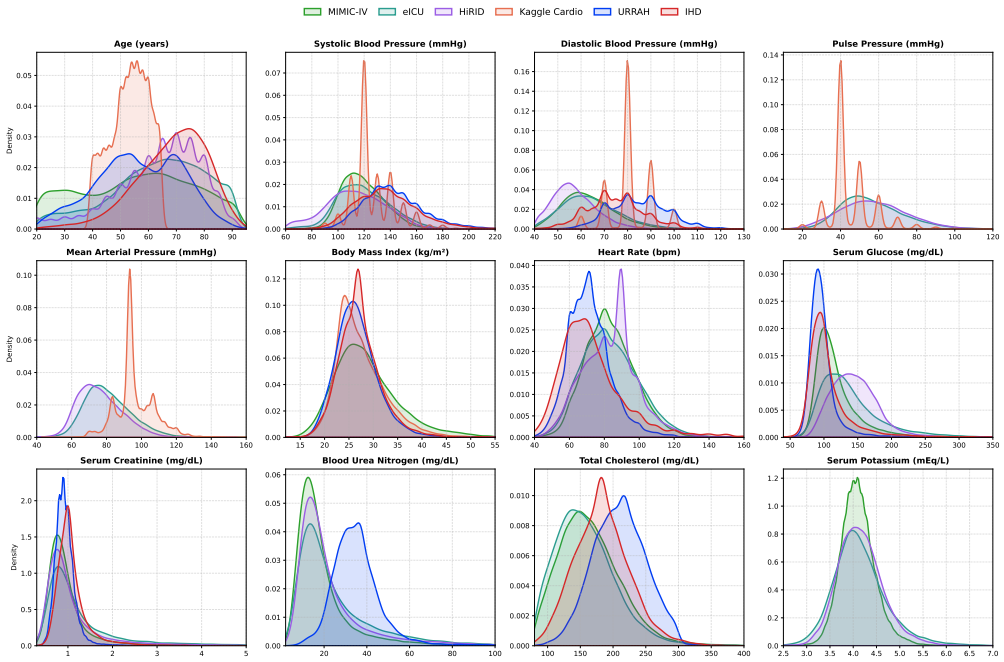


Figure 5: Distributional heterogeneity across clinical cohorts.

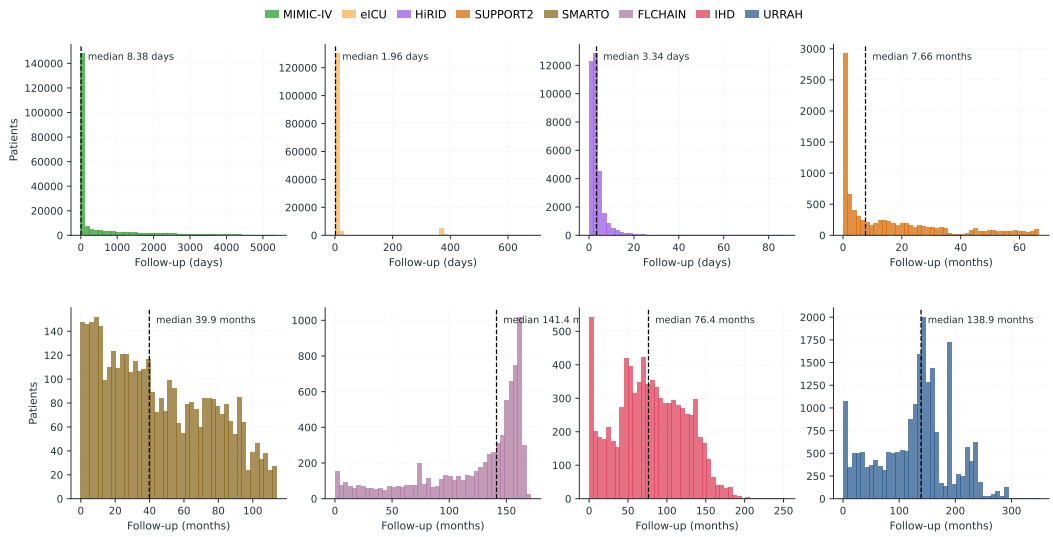


Figure 6: Follow-up time heterogeneity across clinical cohorts.

sparsely observed in a source cohort, and vice versa, while some variables exhibit consistently high or low observability across datasets. Thus, missingness itself undergoes a cross-cohort distribution shift rather than representing a fixed property of a clinical variable. This motivates experiments in Figure 3 by explicitly representing observation state during pre-training and evaluating whether the resulting missingness-aware representation can transfer across cohorts with different observation processes.

Finally, we examine whether exposure to heterogeneous sources changes the geometry of the learned patient representation. Figure 8 compares representations obtained under random initialization, single-source pre-training (on MIMIC-IV), and multi-source pre-training. Under random initialization, the leading principal components exhibit pronounced cohort-specific structure. This sep-



Figure 7: Cross-cohort heterogeneity in feature missingness. Features observed in URRAH are ordered by their target-cohort missingness rate and compared with the corresponding rates in the pre-training cohorts when the same feature is available. Gray cells indicate features not available in a source cohort.

aration is visually reduced after single-source pre-training and further reduced after multi-source pre-training. Correspondingly, the variance explained by the first two principal components decreases from 55.7% under random initialization to 35.0% after MIMIC-IV pre-training and 23.8% after multi-source pre-training, indicating that variation in the representation becomes less concentrated in the leading two-dimensional projection.

Together, these observations motivate the main design choices of SurvBridge. Covariate and schema heterogeneity motivate a shared schema-flexible encoder trained across multiple sources, while differences in follow-up scale motivate the within-source survival coordinate used by SurvVPCL. The representation analysis further shows that after multi-source pre-training, variation is less concentrated in the leading two-dimensional PCA projection and cohort separation is visually less pronounced. These analyses are descriptive and do not imply that cohort-specific distribution shift is eliminated.

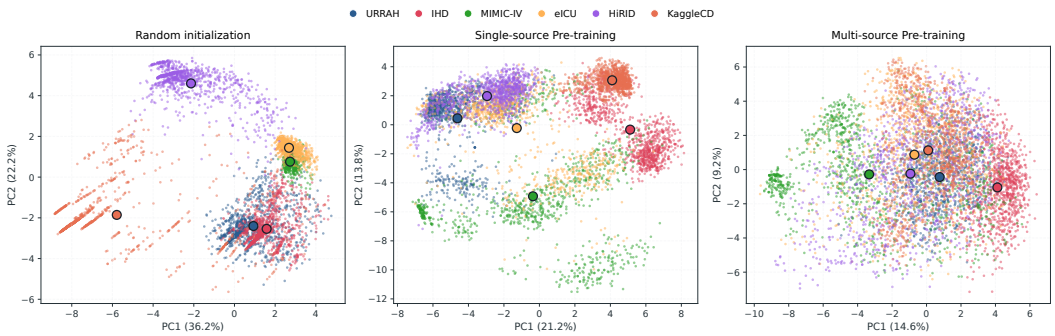


Figure 8: **PCA projection of patient representations across pre-training settings.** Representations from random initialization, single-source pre-training on MIMIC-IV, and multi-source pre-training are projected onto their first two principal components. Colors denote cohorts and outlined markers denote cohort centroids.

A.3 Hyperparameters Tuning

SurvBridge Configuration. All reported SurvBridge experiments use the same architecture: hidden dimension 128, three Transformer layers, eight attention heads, feed-forward dimension 256, projection dimension 128, dropout 0.1, and GELU activation. Contrastive pre-training uses three feature partitions with overlap ratio 0.5, temperature/base temperature 10/10, learning rate 1×10^{-4}, batch size 128, 200 epochs, weight decay 10^{-5}, and patience 20. Survival fine-tuning uses learning rate 10^{-4}, batch size 64, 50 epochs, weight decay 10^{-4}, and patience 10. If training from scratch, we use 200 epochs.

Baseline Defaults and Tuning Spaces. The non-pretrained baselines with model-specific search configurations are tuned independently within each training fold. For baseline models requiring complete inputs, missing numerical values were imputed using the median estimated from the corresponding training partition, with the same estimates applied to validation and test data. Table 3 reports the exact search spaces defined in the experimental configuration files. Continuous parameters marked as log-uniform are sampled on a logarithmic scale; integer ranges are sampled with unit step unless otherwise noted; and categorical parameters are sampled from the listed values.

Prior-Fitted Network (PFN) Models. PFN baselines use frozen pre-trained backbones and are therefore evaluated under their model-specific feature and context constraints rather than Optuna tuning. TabPFN-based survival models use at most 100 covariates and 1,024 training examples as context, whereas SurvivalPFN supports 99 covariates and uses the full training fold as context. Inputs exceeding these feature limits are handled using the preprocessing required by the corresponding implementation. Table 4 summarizes the resulting evaluation protocol.

Alternative Contrastive Pre-training Objectives. For comparison with SurvVPCL in Section 3.4, we additionally pre-train the same encoder using SurvRNC (Saeed et al., 2024) and OTC-Surv (Nayebi Kerdabadi et al., 2023). SurvRNC constructs censoring-aware contrastive relationships from continuous observed follow-up times and is evaluated with temperature $\tau = 2.0$. OTC-Surv instead discretizes follow-up times into nine equal-frequency quantile bins and constructs temporal neighborhoods from the resulting bin indices; we use a neighborhood window of two bins and temperature $\tau = 0.5$. All objectives use the same encoder architecture and multi-source pre-training protocol to provide a controlled comparison of the contrastive objectives.

Table 3: Hyperparameter search spaces for the tuned non-pretrained baselines. Each model is optimized using 30 Optuna trials within each training fold.

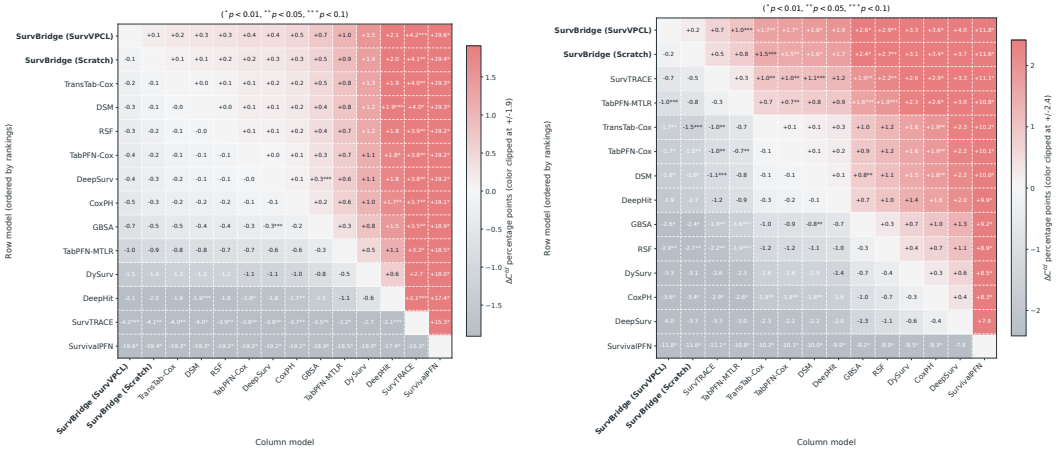
Model	Hyperparameter	Search space
RSF	Number of trees	$[50, 200]$ (integer)
	Minimum samples to split	$[5, 20]$ (integer)
	Minimum samples per leaf	$[5, 20]$ (integer)
GBSA	Learning rate	$[10^{-2}, 2 \times 10^{-1}]$ (log-uniform)
	Number of estimators	$[50, 200]$ (integer)
	Maximum depth	$[2, 5]$ (integer)
DeepHit	Learning rate	$[10^{-4}, 10^{-1}]$ (log-uniform)
	Dropout	$[0, 0.5]$ (uniform)
	Number of hidden layers	$[1, 3]$ (integer)
	Hidden width	$\{32, 64, 128\}$
	DeepHit α	$[0, 0.5]$ (uniform)
	DeepHit σ	$[10^{-2}, 1]$ (log-uniform)
DeepSurv	Learning rate	$[10^{-4}, 5 \times 10^{-2}]$ (log-uniform)
	Dropout	$[0, 0.5]$ (uniform)
	Number of hidden layers	$\{1, 2, 3\}$
	Hidden width	$\{16, 32, 64\}$
	Epochs	$\{10, 20, \ldots, 100\}$
DySurv	Learning rate	$[10^{-4}, 10^{-2}]$ (log-uniform)
	Encoded features	$\{16, 32, 64, 128\}$
	Batch size	$\{64, 128, 256\}$
DSM	Number of mixture components (k)	$\{2, 3, 4, 6\}$
	Distribution	$\{\text{Weibull}, \text{LogNormal}\}$
	Number of hidden layers	$\{1, 2, 3\}$
	Hidden width	$\{50, 100, 150\}$
	Learning rate	$[10^{-4}, 5 \times 10^{-2}]$ (log-uniform)
	Batch size	$\{64, 128, 256\}$
	Training iterations	$\{30, 40, \ldots, 150\}$
SurvTRACE	Learning rate	$[10^{-4}, 10^{-2}]$ (log-uniform)
	Hidden size	$\{8, 16, 32\}$
	Hidden layers	$\{1, 2, 3, 4\}$
	Attention heads	$\{2, 4, 8\}$
	Intermediate size	$\{32, 64, 128\}$

Table 4: Evaluation specifications for prior-fitted network baselines.

Model	Backbone / mode	Input constraints	Survival adaptation
TabPFN-Cox	Frozen TabPFN v1	$D_{\max} = 100$; $N_{\text{ctx}} \leq 1{,}024$	Linear Cox head; Adam, lr $= 10^{-4}$, batch size 64, 50 epochs
TabPFN-MTLR	Frozen TabPFN v1	$D_{\max} = 100$; $N_{\text{ctx}} \leq 1{,}024$	MTLR head (50 bins); Adam, lr $= 10^{-4}$, batch size 64, 50 epochs
SurvivalPFN	Frozen SurvivalPFN; zero-shot	$D_{\max} = 99$; full-fold context	No gradient adaptation; $\tau = 1.0$

A.4 Complete Benchmark Results

Pairwise full-benchmark comparisons. Figure 9 reports paired seed-level differences in C^{td} for the target-native benchmark. On URRAH, all SurvBridge variants outperform most classical, deep, and in-context baselines, while the gap between TransTab and SurvBridge remains fairly large on URRAH. IHD shows the same overall pattern, although differences against the strongest conventional baselines are narrower. These comparisons support the main-text result that SurvBridge is competitive under full target supervision, whereas pre-training provides only a modest additional benefit in the full-data regime.



(a) IHD. (b) URRAH.

Figure 9: Pairwise target-native win matrix across available complete full-feature benchmark models for IHD and URRAH. Cells report paired differences in time-dependent concordance across seeds; positive values favor the row model over the column model, and asterisks denote Holm-adjusted paired tests ($^{*}p < 0.01$, $^{**}p < 0.05$, $^{***}p < 0.1$).

Feature Identity Encoder. Table 5 compares BERT and BioLORD feature identity encoders across three public survival cohorts. BioLORD consistently improves C^{td} over BERT, with paired gains of 0.023, 0.020, and 0.022 on SUPPORT2, SMARTO, and FLCHAIN, respectively. These results suggest that incorporating domain-specific clinical semantics into feature identity representations provides a consistent benefit across heterogeneous survival datasets.

Table 5: Comparison of SurvBridge feature identity encoders on public survival cohorts. Values are mean C^{td} with 95% confidence intervals; Δ denotes the paired difference over matched seeds.

Method	SUPPORT2	SMARTO	FLCHAIN
BERT Devlin et al. (2019)	$0.755_{[0.749,\ 0.761]}$	$0.651_{[0.613,\ 0.689]}$	$0.773_{[0.758,\ 0.788]}$
BioLORD Remy et al. (2022)	$\mathbf{0.778}_{[0.767,\ 0.788]}$	$\mathbf{0.671}_{[0.635,\ 0.707]}$	$\mathbf{0.795}_{[0.780,\ 0.810]}$
Δ (BioLORD $-$ BERT)	$+0.023_{[+0.022,\ +0.024]}$	$+0.020_{[+0.018,\ +0.022]}$	$+0.022_{[+0.020,\ +0.024]}$

Detailed low-resource transfer. Table 6 reports the absolute C^{td} values underlying the paired-gain analysis in Figure 2. Pre-training provides its clearest benefit in the smallest target-data regimes. At $N = 100$, for example, VPCL/SurvVPCL increase mean C^{td} from 0.595 to $0.715/0.712$ on URRAH and from 0.636 to $0.726/0.723$ on IHD. The advantage over training from scratch remains substantial through $N = 200$–500, after which the differences progressively narrow as more target supervision becomes available. VPCL and SurvVPCL exhibit very similar transfer performance across the evaluated sample sizes, with neither uniformly dominating the other. SurvVPCL attains the highest mean at several low-resource settings, including $N = 200$ and 500 on both cohorts, whereas VPCL is slightly higher in others. Both objectives generally provide stronger and more

Table 6: Low-resource transfer detailed results. Rows report Antolini's time-dependent concordance (C^{td}) across matched subsampling seeds as mean and 95% confidence interval. Best and second-best means within each cohort and sample size are **bolded** and underlined, respectively.

URRAH				IHD			
N_{ft}	**Initialization**	**Mean**	**95% CI**	N_{ft}	**Initialization**	**Mean**	**95% CI**
$N = 100$	Scratch	0.595	[0.549, 0.640]	$N = 100$	Scratch	0.636	[0.580, 0.692]
	SurvRnC	0.652	[0.600, 0.703]		SurvRnC	0.701	[0.653, 0.750]
	OTCSurv	0.629	[0.527, 0.731]		OTCSurv	0.645	[0.595, 0.695]
	VPCL	**0.715**	[0.667, 0.764]		VPCL	**0.726**	[0.685, 0.767]
	SurvVPCL	0.712	[0.665, 0.759]		SurvVPCL	0.723	[0.680, 0.766]
$N = 200$	Scratch	0.662	[0.627, 0.697]	$N = 200$	Scratch	0.672	[0.637, 0.708]
	SurvRnC	0.729	[0.699, 0.759]		SurvRnC	0.719	[0.665, 0.773]
	OTCSurv	0.669	[0.629, 0.710]		OTCSurv	0.679	[0.631, 0.726]
	VPCL	0.774	[0.734, 0.814]		VPCL	0.745	[0.707, 0.782]
	SurvVPCL	**0.778**	[0.738, 0.817]		SurvVPCL	**0.751**	[0.708, 0.793]
$N = 500$	Scratch	0.703	[0.659, 0.747]	$N = 500$	Scratch	0.701	[0.685, 0.718]
	SurvRnC	0.765	[0.721, 0.808]		SurvRnC	0.757	[0.744, 0.770]
	OTCSurv	0.720	[0.629, 0.810]		OTCSurv	0.715	[0.662, 0.768]
	VPCL	0.792	[0.757, 0.826]		VPCL	0.771	[0.752, 0.791]
	SurvVPCL	**0.794**	[0.755, 0.834]		SurvVPCL	**0.776**	[0.755, 0.796]
$N = 1,000$	Scratch	0.790	[0.767, 0.813]	$N = 1,000$	Scratch	0.753	[0.724, 0.783]
	SurvRnC	0.804	[0.792, 0.815]		SurvRnC	0.764	[0.745, 0.783]
	OTCSurv	0.784	[0.764, 0.804]		OTCSurv	0.753	[0.690, 0.815]
	VPCL	0.821	[0.811, 0.831]		VPCL	**0.788**	[0.776, 0.800]
	SurvVPCL	**0.822**	[0.816, 0.829]		SurvVPCL	0.783	[0.763, 0.803]
$N = 2,000$	Scratch	0.826	[0.816, 0.836]	$N = 2,000$	Scratch	0.786	[0.773, 0.800]
	SurvRnC	0.827	[0.814, 0.841]		SurvRnC	0.782	[0.763, 0.802]
	OTCSurv	0.817	[0.810, 0.824]		OTCSurv	0.774	[0.733, 0.814]
	VPCL	**0.842**	[0.835, 0.849]		VPCL	**0.794**	[0.779, 0.808]
	SurvVPCL	0.835	[0.826, 0.844]		SurvVPCL	0.793	[0.776, 0.810]
$N = 4,000$	Scratch	0.839	[0.832, 0.845]	$N = 4,000$	Scratch	0.795	[0.786, 0.805]
	SurvRnC	0.842	[0.837, 0.847]		SurvRnC	0.792	[0.778, 0.805]
	OTCSurv	0.835	[0.825, 0.844]		OTCSurv	0.798	[0.791, 0.805]
	VPCL	**0.848**	[0.840, 0.856]		VPCL	0.799	[0.790, 0.808]
	SurvVPCL	0.847	[0.840, 0.854]		SurvVPCL	**0.800**	[0.785, 0.815]
$N = 6,000$	Scratch	0.847	[0.840, 0.854]				
	SurvRnC	0.849	[0.841, 0.858]				
	OTCSurv	0.839	[0.829, 0.849]				
	VPCL	**0.852**	[0.846, 0.858]				
	SurvVPCL	0.850	[0.841, 0.860]				
$N = 10,000$	Scratch	0.850	[0.843, 0.857]				
	SurvRnC	0.855	[0.849, 0.862]				
	OTCSurv	0.846	[0.837, 0.855]				
	VPCL	0.858	[0.851, 0.866]				
	SurvVPCL	**0.859**	[0.854, 0.865]				

persistent low-resource transfer than SurvRNC and OTCSurv. SurvRNC also yields substantial gains in several small-sample settings, while OTCSurv provides smaller and more variable improvements. Overall, the results indicate that preserving VPCL's cross-view self-supervised pre-training structure is particularly effective for cross-cohort transfer, while survival weighting can incorporate source outcome information without sacrificing this transfer behavior.

Large-EHR survival benchmark. As an additional evaluation of the survival architecture, we construct separate downstream survival tasks from MIMIC-IV and eICU, denoted MIMIC-IV-Surv and eICU-Surv. These downstream cohorts are constructed separately from the corresponding pre-training partitions, and train/test separation is preserved to prevent patient- or outcome-level leakage.

Table 7 reports discrimination performance (C^{td}) on these datasets. SurvBridge achieves top performance on eICU-Surv while remaining competitive to foundation model on MIMIC-IV-Surv. These

results provide comprehensive external validation of the survival architecture, complementary to the cross-cohort transfer experiments that form the main focus of the paper.

Table 7: Predictive survival performance on the acute-care EHR cohorts, measured by time-dependent concordance index (C^{td} ↑) and integrated Brier score (IBS ↓). Mean and 95% confidence intervals are computed across seeds. Best and second-best means within each dataset and metric are **bolded** and underlined, respectively.

Model	MIMIC-IV-Surv		eICU-Surv	
	C^{td} ↑	IBS ↓	C^{td} ↑	IBS ↓
Classical survival models				
CoxPH	$0.787_{[0.770,\,0.804]}$	$0.248_{[0.189,\,0.308]}$	$0.717_{[0.692,\,0.743]}$	$0.177_{[0.163,\,0.192]}$
RSF	$0.801_{[0.781,\,0.821]}$	$0.241_{[0.202,\,0.280]}$	$0.720_{[0.702,\,0.738]}$	$\underline{0.171}_{[0.160,\,0.182]}$
GBSA	$0.787_{[0.767,\,0.806]}$	$0.237_{[0.182,\,0.291]}$	$0.709_{[0.690,\,0.729]}$	$0.178_{[0.165,\,0.191]}$
Deep survival models				
DeepSurv	$0.627_{[0.472,\,0.781]}$	$0.264_{[0.210,\,0.318]}$	$0.717_{[0.631,\,0.803]}$	$0.179_{[0.164,\,0.194]}$
DeepHit	$0.810_{[0.807,\,0.812]}$	$0.326_{[0.318,\,0.334]}$	$0.733_{[0.728,\,0.737]}$	$0.191_{[0.189,\,0.193]}$
DySurv	$0.839_{[0.827,\,0.852]}$	$0.358_{[0.269,\,0.447]}$	$0.772_{[0.761,\,0.783]}$	$0.459_{[0.404,\,0.514]}$
SurvTRACE	$0.775_{[0.756,\,0.794]}$	$\mathbf{0.217}_{[0.171,\,0.263]}$	$0.781_{[0.772,\,0.790]}$	$\mathbf{0.169}_{[0.155,\,0.184]}$
DSM	$0.838_{[0.836,\,0.840]}$	$\underline{0.226}_{[0.222,\,0.229]}$	$0.783_{[0.781,\,0.785]}$	$0.174_{[0.172,\,0.176]}$
TransTab-Cox	$0.834_{[0.818,\,0.850]}$	$0.295_{[0.204,\,0.387]}$	$0.786_{[0.778,\,0.793]}$	$0.182_{[0.166,\,0.198]}$
Tabular foundation models				
SurvivalPFN	$0.597_{[0.594,\,0.601]}$	$0.271_{[0.270,\,0.272]}$	$0.637_{[0.636,\,0.637]}$	$0.199_{[0.199,\,0.199]}$
TabPFN-Cox	$\underline{0.844}_{[0.834,\,0.854]}$	$0.274_{[0.201,\,0.346]}$	$0.779_{[0.772,\,0.785]}$	$0.183_{[0.167,\,0.199]}$
TabPFN-MTLR	$\mathbf{0.845}_{[0.836,\,0.853]}$	$0.329_{[0.240,\,0.418]}$	$0.783_{[0.776,\,0.789]}$	$0.260_{[0.228,\,0.292]}$
SurvBridge (Ours)				
SurvBridge (*Scratch*)	$0.841_{[0.831,\,0.851]}$	$0.272_{[0.203,\,0.336]}$	$\underline{0.788}_{[0.781,\,0.795]}$	$0.185_{[0.171,\,0.198]}$
SurvBridge (*SurvVPCL*)	$\underline{0.844}_{[0.834,\,0.853]}$	$0.270_{[0.196,\,0.348]}$	$\mathbf{0.790}_{[0.785,\,0.795]}$	$0.183_{[0.170,\,0.196]}$

Table 8: Comparison of Multi-source (Large-scale EHRs, SurvSet-Clinic), and MIMIC single-source pretrained models vs. training from scratch. Metrics report mean ± std across 5 random seeds with 5-fold cross-validation. Best results per dataset are in **bold**.

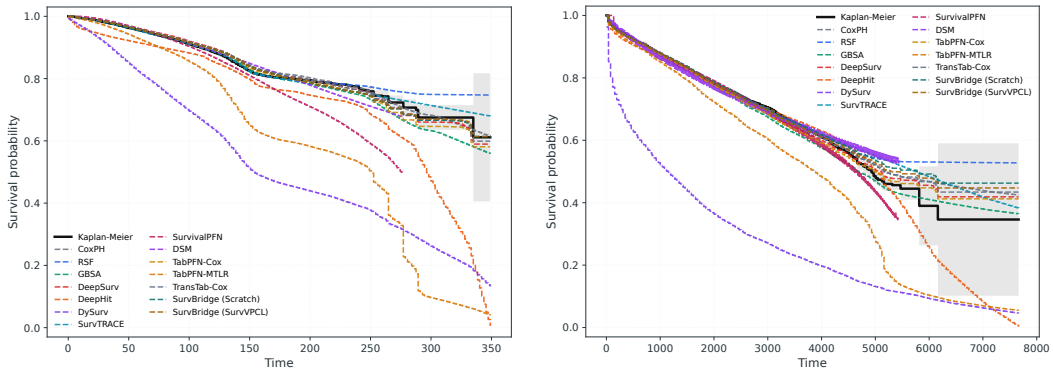
Dataset	Metric	Multi-source		Single-source	Scratch
		Large-scale EHRs	SurvSet-Clinic	MIMIC-IV	
FLCHAIN	C^{td} (↑)	**0.795**$_{\pm 0.012}$	0.793$_{\pm 0.012}$	0.794$_{\pm 0.012}$	0.773$_{\pm 0.012}$
	Convergence Time (↓)	**405 s**	693 s	452 s	810 s
SMARTO	C^{td} (↑)	**0.671**$_{\pm 0.030}$	0.662$_{\pm 0.023}$	0.658$_{\pm 0.029}$	0.651$_{\pm 0.031}$
	Convergence Time (↓)	**179 s**	262 s	197 s	425 s
SUPPORT2	C^{td} (↑)	**0.778**$_{\pm 0.007}$	**0.778**$_{\pm 0.007}$	0.777$_{\pm 0.007}$	0.755$_{\pm 0.005}$
	Convergence Time (↓)	**689 s**	1580 s	981 s	1724 s
IHD	C^{td} (↑)	**0.802**$_{\pm 0.011}$	0.796$_{\pm 0.009}$	0.801$_{\pm 0.007}$	0.800$_{\pm 0.008}$
	Convergence Time (↓)	**375 s**	741 s	552 s	1001 s
URRAH	C^{td} (↑)	0.859$_{\pm 0.005}$	**0.861**$_{\pm 0.005}$	0.857$_{\pm 0.005}$	0.855$_{\pm 0.005}$
	Convergence Time (↓)	**1348 s**	3006 s	2076 s	2671 s
Summary (Macro Avg)	C^{td} (↑)	**0.781**$_{\pm 0.007}$	0.778$_{\pm 0.005}$	0.777$_{\pm 0.006}$	0.767$_{\pm 0.006}$
	Convergence Time (↓)	**599 s**	1257 s	852 s	1326 s

Pre-training Accelerates Target Adaptation. Beyond predictive performance, we examine whether pre-training provides a more efficient initialization for downstream survival adaptation. In addition to our multi-source large-EHR pre-training, we consider two alternative pre-training sources: MIMIC-IV alone and **SurvSet-Clinic**, a curated collection of 34 open-access clinical survival datasets from SurvSet comprising 35,081 patients across diverse clinical domains. All SurvSet-

Clinic cohorts provide time-to-event labels and retain their native clinical feature spaces. Evaluation targets and cohorts with potential provenance overlap are excluded from the collection.

Table 8 compares these initialization strategies with training SurvBridge from scratch. Multi-source large-EHR pre-training converges faster than training from scratch on all five target cohorts, reducing the macro-average target adaptation time from 1326 s to 599 s (55%). It also converges faster than either single-source MIMIC-IV or SurvSet-Clinic pre-training while achieving the highest macro-average C^{td}. These results indicate that multi-source pre-training provides not only a transferable representation but also an effective initialization for efficient downstream survival adaptation.

Population-Level Survival Calibration. Figure 10 compares the empirical Kaplan–Meier curve with the mean predicted survival function of each model on the held-out target cohorts. Most models broadly follow the empirical trajectory during early and intermediate follow-up, although several show larger deviations later in time. SurvBridge remains comparatively close to the Kaplan–Meier trajectory across much of the observed follow-up in both cohorts.



(a) URRAH (b) IHD

Figure 10: **Observed and predicted population survival curves.** Kaplan–Meier estimates are compared with the mean predicted survival curves of each model on the held-out test sets.

Risk-stratified survival trajectories. Figure 11 visualizes predicted survival trajectories after grouping patients into low-, medium-, and high-risk strata according to each model's own predictions. All three models produce an ordered separation between risk groups on both URRAH and IHD. For SurvBridge, the predicted trajectories show a persistent risk gradient over follow-up: the high-risk group exhibits the steepest decline in predicted survival, the low-risk group remains comparatively stable, and the medium-risk group follows an intermediate trajectory. RSF shows the same ordering with narrower separation, while TabPFN-MTLR produces more rapidly declining survival estimates at later follow-up times. Because the strata are defined from each model's own predictions, these plots are interpreted as a qualitative diagnostic of risk separation rather than as independent evidence of calibration.

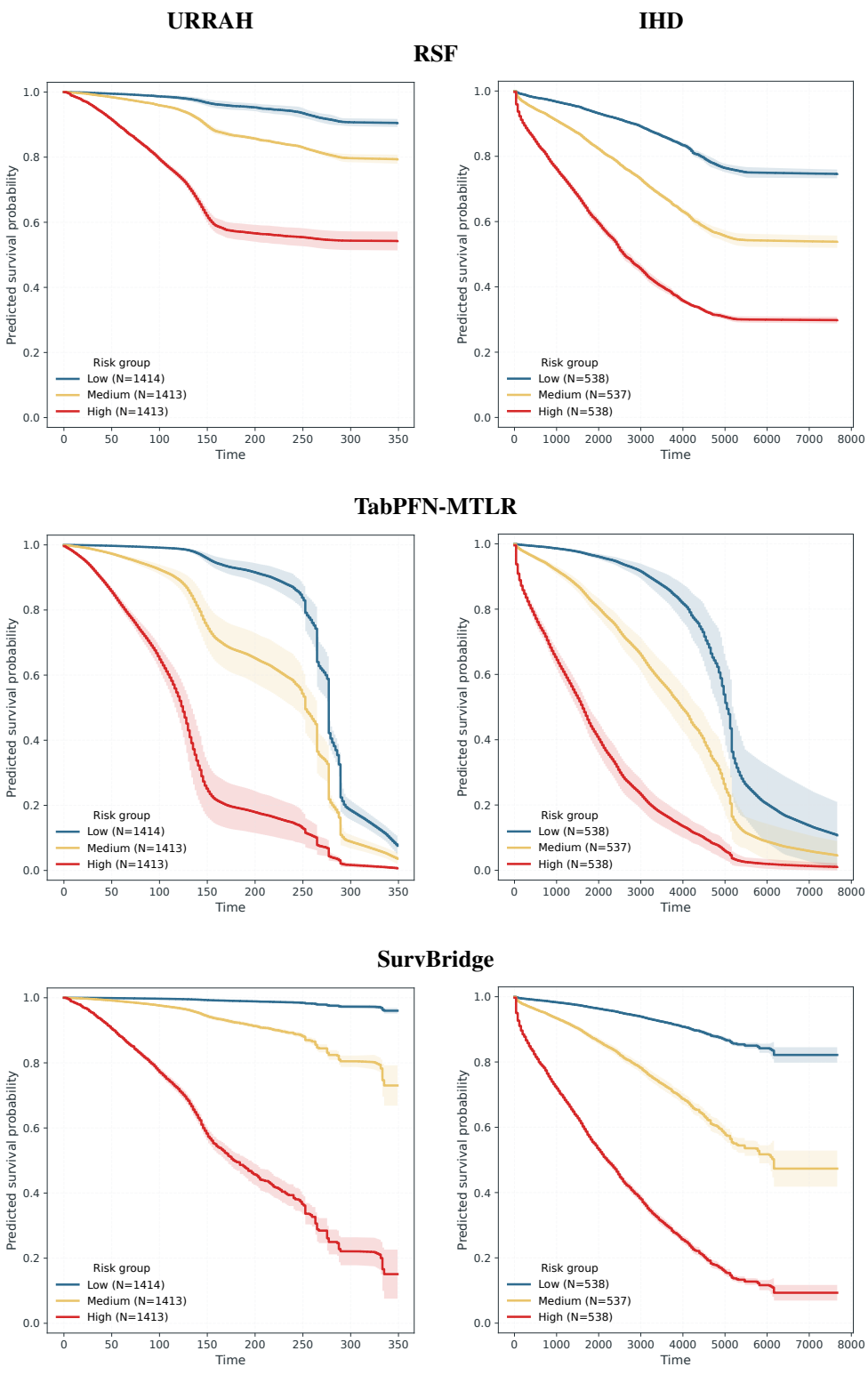


Figure 11: Risk-stratified predicted survival trajectories for baselines, and SurvBridge on URRAH and IHD. Patients are grouped into low-, medium-, and high-risk categories according to each model's predicted risk. Shaded regions summarize variation in predicted survival across patients within each risk stratum. The figure is intended as a qualitative assessment of risk separation rather than an independent calibration evaluation.