The Severity Trap: Benchmarking Sequential Representation Learning for Oxygenation Trajectory Phenotyping in Acute Hypoxaemic Respiratory Failure

Dominic C. Marshall DOMINIC.MARSHALL12@IMPERIAL.AC.UK
Nikita Narayanan NIKITA.NARAYANAN18@IMPERIAL.AC.UK
David B. Antcliffe D.ANTCLIFFE@IMPERIAL.AC.UK
Sonali Parbhoo S.PARBHOO@IMPERIAL.AC.UK
Imperial College London, London, United Kingdom

Abstract

Patients with acute hypoxaemic respiratory failure (AHRF) presenting with similar early physiology can follow markedly different trajectories, motivating trajectory phenotyping for mechanistic discovery and trial enrichment. Statistical models such as competing-risk latent class mixed models identify clinically meaningful classes but scale poorly to high-dimensional, irregular ICU time series, motivating sequential representation learning with variational autoencoders (VAEs). We characterise a systematic failure mode of learned trajectory representations, the *severity trap*: representations organise primarily along baseline illness severity rather than trajectory shape, even with auxiliary survival objectives or disentanglement constraints. This is an instance of shortcut learning, or nuisance-variable dominance, that we characterise for clinical trajectory phenotyping and show evades the unsupervised metrics commonly used to select such models. Using engineered baselines of simple trajectory features that match, and once given the same outcome signal exceed, every deep model, together with controlled synthetic experiments, we trace the failure to reconstruction-based objectives preferentially encoding the highest-variance factor of the input, here baseline severity. We introduce a multi-cohort benchmark evaluating sequential representation learners (VAE families as the primary object of study; masked-transformer, diffusion, and contrastive encoders as stress tests) against four externally validated oxygenation trajectory archetypes across three international ICU cohorts (MIMIC-IV, Imperial College Healthcare NHS Trust, and Amsterdam UMCdb). Every reconstruction- or level-preserving objective falls into the trap. Only an explicitly level-invariant contrastive objective partially escapes, and clustering stability and reconstruction error correlate little with clinical validity. These findings highlight the need for evaluation grounded in externally validated clinical phenotypes, with nuisance-aware baselines, when developing representation learning for healthcare time series.

1. Introduction

Acute hypoxaemic respiratory failure (AHRF) is a common ICU syndrome with substantial mortality (Pham et al., 2021; Ling et al., 2024). Patients with similar initial physiology can follow divergent longitudinal respiratory courses (Loewen et al., 2024; Chen et al., 2022), yet most subphenotyping work in ARDS remains cross-sectional, assigning patients to classes at a single early time-point (Calfee et al., 2014; Sinha et al., 2018). As hypoxaemia persists, the initial insult may resolve or be compounded by ICU-acquired "second hits," such that longitudinal phenotyping is needed to reduce stage-mixing and scaffold mechanistic

discovery. Recent work using a competing-risk latent class mixed model (CRLCMM) identified four AHRF trajectory archetypes: early recovery, stable persistence, biphasic, and rapid decline. These share similar baseline oxygenation yet diverge in temporal course and outcome (Marshall et al., 2026). Extending such trajectory models to irregularly sampled ICU time series remains challenging.

Sequential representation learning, particularly with variational autoencoders (VAEs) , but also contrastive, masked-reconstruction, and diffusion encoders, has been proposed for learning latent representations from irregular clinical time series (Kingma and Welling, 2013; Che et al., 2018; Shukla and Marlin, 2021; Rubanova et al., 2019). We take VAEs as our primary object of study and use the oxygenation trajectory identification task as a controlled test case, with non-generative and non-variational encoders as stress tests. These approaches implicitly assume that accurate reconstruction and stable latent clusters correspond to clinically interpretable structure. We show that this assumption can be misleading in trajectory phenotyping: representations that achieve excellent reconstruction and clustering stability may primarily encode baseline illness severity rather than the temporal dynamics that distinguish clinically meaningful trajectories. We term this failure mode the *severity trap*, and identify it as an instance of shortcut learning (a model exploiting a dominant, easily encoded nuisance factor) in the specific setting of clinical trajectory phenotyping (Geirhos et al., 2020).

Contributions. i) We characterise the *severity trap* in clinical trajectory phenotyping, an instance of shortcut learning / nuisance-variable dominance, in which sequential representation learners organise along baseline severity rather than trajectory shape. We also provide a simple diagnostic (a severity probe) and a rate-distortion account of why reconstruction-based objectives do this. ii) Through controlled synthetic experiments we show the failure is driven by the variance structure of the data, not by architecture: the identical model recovers trajectory shape once shape, rather than severity, dominates input variance. iii) We show that the unsupervised metrics used to select these models (reconstruction error and clustering stability) are poor proxies for clinical validity, and that trajectory shape is linearly decodable from every latent yet is never recovered by unsupervised clustering, indicating a geometric rather than a capacity failure. iv) We introduce a multi-cohort benchmark evaluating sequential representation learners, with VAE families as the primary object of study and masked-transformer, diffusion, and contrastive encoders as stress tests, against externally validated trajectory phenotypes across three international ICU cohorts, and quantify the clinical cost of the trap for trial enrichment.

Generalizable Insights about Machine Learning in the Context of Healthcare

The Severity Trap. Sequential representation learners, VAEs, and more generally reconstruction or level-preserving objectives can produce latent representations that are informative and stable yet systematically aligned with baseline illness severity rather than the temporal dynamics of clinical interest. This is an instance of shortcut learning, or nuisance-variable dominance (Geirhos et al., 2020), in clinical trajectory phenotyping; standard unsupervised metrics fail to detect the misalignment. More broadly, a dominant static severity or acuity axis is common in clinical time series (sepsis, acute kidney injury, and post-operative

deterioration share this structure), so any unsupervised representation of such data risks encoding where a patient starts rather than where they are heading; the lesson generalises to any healthcare setting where a high-variance baseline nuisance co-varies with the dynamic of clinical interest.

Auxiliary objectives mitigate but do not guarantee disentanglement. Outcome prediction tasks can partially encourage prognostically relevant gradients, but do not necessarily separate latent factors that share similar outcomes yet differ in temporal dynamics. This matters because outcome-supervised representation learning is a common solution in clinical ML, yet states that share an outcome but differ in tempo, precisely the cases where the timing of treatment matters, can remain entangled; auxiliary supervision should be validated against the dynamic of interest rather than assumed to disentangle it.

Domain-informed baselines remain competitive. Simple engineered features capturing trajectory shape are competitive with deep representation learning models, and exceed them when given the same survival signal the deep models receive, highlighting the importance of domain-informed baselines and nuisance-aware evaluation when assessing ML methods for healthcare time series.

2. Related Work

AHRF/ARDS subphenotyping. Latent class analyses of cross-sectional clinical and biomarker data reproducibly identify hyper- and hypo-inflammatory ARDS subphenotypes (Calfee et al., 2014; Sinha et al., 2018; Famous et al., 2017), but assign phenotypes at a single early time-point. Longitudinal extensions using multivariate clinical data (Chen et al., 2022) or PaO_2/FiO_2 trajectories (Loewen et al., 2024) show that trajectory shape carries prognostic information beyond static severity; we evaluate whether deep representation learning can recover such structure from ICU time series.

Trajectory clustering in critical care. Group-based trajectory modelling has been applied to lactate (Wang et al., 2022), mechanical power (Song et al., 2025), vital signs (Li et al., 2024), and respiratory mechanics in COVID-19 ARDS (Bos et al., 2021), but these approaches typically operate over short windows and do not account for informative censoring. The CRLCMM (Marshall et al., 2026) jointly models trajectories and competing risks, enabling discovery of non-monotonic trajectory classes. We use these archetypes as a reference structure for evaluating deep models.

Deep representation learning for clinical time series. VAEs (Kingma and Welling, 2013) with missingness-aware encoders (GRU-D (Che et al., 2018)), attention mechanisms (mTAN (Shukla and Marlin, 2021)), and continuous-time dynamics (Latent ODEs (Rubanova et al., 2019)) have been proposed for learning compact trajectory representations. We provide a systematic evaluation of whether such methods recover interpretable trajectory structure.

Shortcut learning, nuisance variables, and VAE failure modes. Deep models frequently exploit *shortcuts*: easily encoded features that reduce training loss but do not correspond to the intended structure (Geirhos et al., 2020), a nuisance-variable dominance in which a high-variance factor crowds out the factor of interest. The severity trap is a clinically consequential instance: baseline illness severity is the dominant, easily reconstructed factor and crowds out trajectory shape. It is distinct from posterior collapse, an ELBO

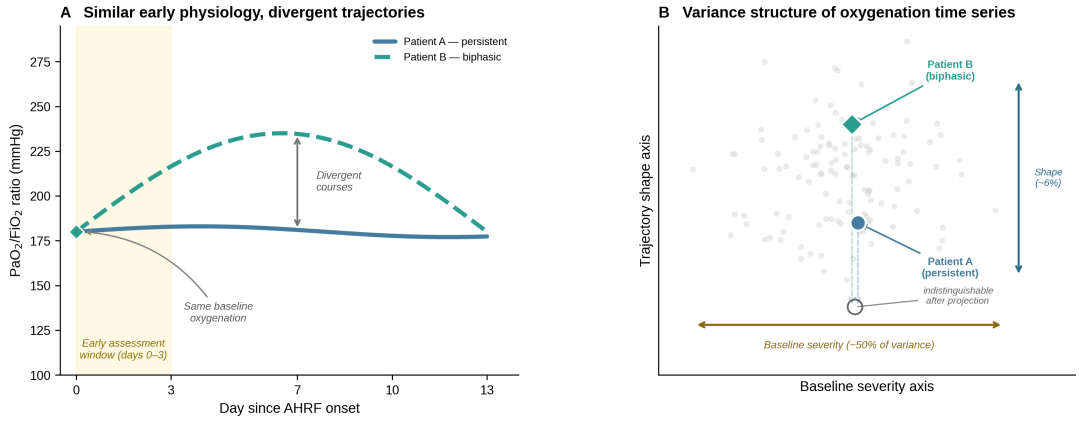


Figure 1: The severity trap. **(A)** Two AHRF patients with near-identical early oxygenation follow divergent later courses (one persistent, one biphasic). **(B)** In oxygenation time series, baseline severity dominates the variance while trajectory shape contributes far less; a representation that compresses along the high-variance severity axis renders differently shaped trajectories nearly indistinguishable.

pathology where latent variables become uninformative (Bowman et al., 2016; Chen et al., 2017; Lucas et al., 2019; Alemi et al., 2018): here the representation stays informative and clusterable but organises along the nuisance axis rather than the temporal structure trajectory phenotyping requires. We characterise this in clinical trajectory phenotyping, show the standard unsupervised selection metrics miss it, and benchmark it across architectures and cohorts.

3. The Severity Trap in Deep Trajectory Phenotyping

We define the *severity trap* as a failure mode in which ELBO-trained sequential VAEs learn latent representations that organise primarily along baseline illness severity rather than the temporal trajectory dynamics that define distinct phenotypes. This occurs because baseline severity dominates the variance of oxygenation time series (∼50%; Appendix I), and the reconstruction objective preferentially encodes the signal that most reduces reconstruction loss. Patients with identical baseline severity but different trajectory shapes are not indistinguishable to a supervised linear classifier, but they are poorly separated according to the latent geometry used by unsupervised clustering. Figure 1 illustrates this problem.

3.1. Problem Setup.

We study trajectory phenotyping in patients with AHRF. Each ICU stay is represented by a daily oxygenation trajectory over a fixed 14-day observation window beginning at AHRF onset. Formally, for each patient ICU stay, we observe a time series of $\mathrm{PaO_2/FiO_2}$ (P/F) measurements $\mathbf{x} = (x_1, \ldots, x_{14})$ with observation mask $\mathbf{m} = (m_1, \ldots, m_{14})$ where $m_t = 1$ on day t if the P/F ratio is available and 0 otherwise. Given a dataset of N ICU stays, $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{m}_i)\}_{i=1}^{N}$, our goal is to learn a representation of trajectories that captures

clinically meaningful patterns of temporal dynamics. In particular, we seek representations that distinguish trajectories with similar baseline severity but different temporal evolution.

Deep Phenotyping Objective. Standard deep trajectory phenotyping approaches employ variational latent-variable models to learn representations of trajectory phenotypes (Kingma and Welling, 2013; Rubanova et al., 2019). A VAE encoder $q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{m})$ maps each stay to a latent embedding $\mathbf{z} \in \mathbb{R}^d$, and a decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$ reconstructs the trajectory. Models are trained by maximising the evidence lower bound:

$$\mathcal{L} = \mathbb{E}_{q_\phi}\big[\log p_\theta(\mathbf{x} \mid \mathbf{z})\big] - \beta\,\mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\big) - \lambda_s\,\ell_{\mathrm{surv}}(\mathbf{z}) \tag{1}$$

where β controls the KL weight (with linear warmup), ℓ_{surv} is an optional auxiliary survival loss, and λ_s its weight (λ_s=0 for reconstruction-only models). Reconstruction loss is computed only on observed time points (masked MSE). For notational brevity, we write $q_\phi(\mathbf{z} \mid \mathbf{x})$ below, leaving the conditioning on $\mathbf{m}$ implicit.

Definition 1 (Severity Trap) *A learned representation $z \sim q_\phi(z \mid x)$ exhibits the severity trap if $I(z; s(x)) \gg I(z; \tau(x))$, where $s(x)$ denotes baseline severity and $\tau(x)$ trajectory-shape features.*

This failure mode is distinct from both posterior collapse (Alemi et al., 2018), since the representation remains informative, and from the limits of unsupervised disentanglement (Locatello et al., 2019). Instead, it reflects how the ELBO allocates limited capacity toward the factor that most reduces reconstruction error. A theoretical analysis is provided in Appendix A.

In what follows, we perform systematic analyses on test-set embeddings, clustering them after training and comparing to the reference trajectories from (Marshall et al., 2026).

Reference Standard Trajectory Classes. The reference trajectories in (Marshall et al., 2026), denoted $\mathrm{TC}_{\mathrm{ref}}$, were derived from a competing-risk latent class mixed model (CRLCMM) fitted to the MIMIC-IV cohort. The CRLCMM jointly models longitudinal P/F ratio trajectories via class-specific quadratic polynomials in time (with a random intercept and slope) and competing risks of ICU death versus discharge via class-specific proportional hazards. The four classes (Table 1) span qualitatively distinct trajectory shapes: early recovery (TC1), stable persistence (TC2), biphasic improvement-then-deterioration (TC3), and rapid decline (TC4). Critically, TC1-4 share similar baseline oxygenation and comorbidity burden yet diverge in trajectory and outcome: evidence that these classes capture temporal dynamics beyond static severity. The CRLCMM transfers to both external cohorts with high assignment certainty (median posterior probability: EUKC 0.95, ENC 0.84) and reproduces qualitatively similar trajectory profiles and outcome curves at each site (Marshall et al., 2026). Additional details on the reference standard trajectory classes are provided in Appendix G; the full methodology is published in *Intensive Care Medicine* (Marshall et al., 2026).

3.2. Synthetic Demonstration of the Severity Trap

Setup. We construct a synthetic cohort of N=1,600 patient trajectories each belonging to one of 4 possible phenotype classes corresponding to early recovery, persistent AHRF, initial improvement followed by deterioration (biphasic), and steady deterioration respectively as

Class	Description	N (%)	14-d Mort.	Trajectory
TC1	Early recovery	764 (19.4%)	0.3%	Improving, early discharge
TC2	Stable persistence	1,668 (42.4%)	8.0%	Flat / slowly improving
TC3	Biphasic	754 (19.1%)	17.0%	Improve → deteriorate
TC4	Rapid decline	752 (19.1%)	100.0%	Monotonically worsening

Table 1: TC_{ref} classes in MIMIC-IV (N=3,938). Mortality is 14-day ICU mortality (Marshall et al., 2026). TC4's 100% 14-day mortality reflects the model's competing-risk structure: TC4 defines the rapidly deteriorating stratum.

in (Marshall et al., 2026). Specifically, trajectories belonging to classes TC2 and TC3 are constructed such that they share an identical baseline P/F distribution (μ=175 mmHg, σ=28 mmHg). This ensures that any difference in representation must be driven by trajectory shape alone. TC1-like trajectories start higher (μ=225 mmHg) and improve; TC4-like start lower (μ=155 mmHg) and decline. Day-level noise (σ=9 mmHg) and ∼30% random missingness match MIMIC-IV characteristics. Full generation details are in Appendix B.

We compare two VAE models, Model A (reconstruction-only, trained from scratch) and Model B (MIMIC-pretrained with survival auxiliary, frozen encoder), against engineered baselines using GMM clustering (K=4).

Results. Both VAE models fail to separate TC2 and TC3 (Table 4 and Fig. 4, Appendix B). The frozen Model B encoder organises trajectories along a single severity axis (PC1 explains 78.9% of variance), with TC2 and TC3 completely interleaved (ARI = 0.024, random). Model A, trained from scratch on synthetic data with no MIMIC exposure, produces nearly identical failure (TC2/TC3 ARI = 0.047), confirming that the severity trap arises from the reconstruction objective itself rather than dataset-specific biases. In contrast, engineered slope features separate TC2 from TC3 in this synthetic setting where the two classes differ *only* in shape, demonstrating that the trajectory signal is present in the data but not captured by the learned representations.

The trap is variance-structure-driven, not architectural. To separate an architectural failure from a property of the data, we invert the variance ratio: a *shape-dominant* synthetic cohort in which all four classes share the same baseline (μ=175 mmHg, σ=8 mmHg) while shape amplitudes are large (TC1 +90, TC3 ±90, TC4 −12/day). The identical Model A, retrained from scratch on a separately generated cohort, now *recovers* trajectory shape: TC2/TC3 ARI rises from 0.036 in the severity-dominant arm to 0.770 in the shape-dominant arm, and severity leakage falls from R^2=0.391 to R^2≈0. (The severity-dominant 0.036 here and the 0.047 for this model in Table 4 are counterparts from separate synthetic draws; both are near-random.) The structured decoders (D1–D3, defined in Section 5) behave the same way (TC2/TC3 ARI): near-zero on the severity-dominant cohort (D1 0.011, D2 0.000, D3 0.002) yet high on the shape-dominant cohort (0.879, 0.765, 0.857), so the decoders do not fix the trap; the variance ratio does. A sample-size sweep excludes undertraining as the cause: Model A's TC2/TC3 ARI stays $\leq$ 0.13 from N=1,600 to N=50,000 under severity dominance, versus 0.77 under shape dominance. The VAE encodes whichever factor dominates reconstruction variance, rather than exhibiting a fixed architectural preference.

4. Cohort Selection and Data

We study adults (≥18 years) admitted to ICUs in the MIMIC-IV database (version 1.0) (Johnson et al., 2023, 2021) who met the following criteria for persistent AHRF: PaO_2/FiO_2 (P/F) ratio <300 mmHg with PEEP ≥5 cmH_2O sustained for ≥72 hours, or ≥48 hours if death occurred on day 3. AHRF onset (day 0) is defined as the occasion when both oxygenation and ventilation criteria were met. We extract daily mean P/F ratio over days 0–13 (14-day window), yielding a univariate time series per ICU stay. Patient stays are split 70/10/20 into train, validation, and test sets stratified by ICU mortality (N_{total}=3,938; N_{train}=2,756, N_{val}=394, N_{test}=788). Event outcomes (ICU death and ICU discharge, recorded daily) are used as targets for the survival auxiliary loss (Section 5.2) but not as model inputs; the model input is exclusively the univariate P/F time series and its observation mask. The main findings replicate on the current MIMIC-IV v3.1 release (Appendix N).

Missingness. Daily P/F ratios were unavailable on 34.2% of patient-days (median 9 observed days per stay). Missing values were not imputed for deep models; instead, each model received a binary observation mask alongside the P/F values, and temporal encoders (GRU-D and BiGRU) used the mask to modulate hidden-state updates.

External cohorts. Two additional external AHRF cohorts are used for validation: a UK cohort from Imperial College Healthcare NHS Trust (EUKC; 3 tertiary ICUs, London, UK; N=2,888) and the Netherlands cohort from AmsterdamUMCdb (Thoral et al., 2021) (ENC; 1 tertiary ICU; N=3,592). Both cohorts applied the same AHRF inclusion criteria and 14-day observation window. P/F ratios in EUKC were recorded in kPa and converted to mmHg before model application. Daily P/F is observed on ≈67-77% of the 14-day grid in the external cohorts (versus ≈66% in MIMIC), with no imputation; the trap therefore reproduces externally under *lower* missingness than in the derivation cohort. Cohort characteristics are summarised in Appendix J (Table 9); the three cohorts have comparable demographics and respiratory parameters.

5. Experimental Setup

5.1. Baseline Model Architectures and Training Details

We evaluate seven architectures spanning the main design axes explored in prior work on clinical time-series representation learning: encoder family, decoder structure, auxiliary survival objectives, and latent space organisation (summarised in Table 5, Appendix D). Hyperparameters were tuned independently for each architecture.

Models A (GRU-D / BiGRU): reconstruction-only baselines. These models implement standard sequential VAEs trained only with the ELBO objective. A GRU-D or bidirectional GRU encoder maps the irregularly sampled sequence to a latent posterior $q(z|\mathbf{x})$, and a GRU autoregressive decoder reconstructs the trajectory. These models serve as reconstruction-only baselines without auxiliary objectives or structured latent spaces.

Model B: survival-auxiliary SeqVAE. Model B augments the GRU-D architecture with a discrete competing-risk survival classifier (See Section 5.2). The survival loss is added to

the ELBO objective to encourage the latent representation to capture prognostic information relevant to patient outcomes.

Model C: unsupervised split-latent VAE. Model C introduces a factorised latent representation consisting of a *level* subspace ($z_\ell \in \mathbb{R}^8$) intended to capture baseline severity and a *shape* subspace ($z_s \in \mathbb{R}^8$) intended to capture trajectory dynamics. Auxiliary regularisation encourages the two subspaces to represent distinct factors of variation.

Models D1–D3: structured trajectory decoders. The final family of models incorporates structural assumptions about trajectory dynamics while retaining the survival auxiliary objective. D1 replaces the autoregressive decoder with a learnable basis-function decoder that encourages smooth trajectory ($n = 5$ learnable basis vectors). D2 introduces a linear state-space model (SSM), which applies a learnable transition matrix to z_s at each time step, promoting structured dynamics. D3 augments the same SSM decoder with two disentanglement mechanisms. First, a cross-covariance orthogonality penalty discourages correlation between the level and shape subspaces,

$$\mathcal{L}_\perp = \lambda_\perp \left\| \mathrm{Cov}(\mu_\ell, \mu_s) \right\|_F^2, \qquad \lambda_\perp = 0.1, \tag{2}$$

where μ_ℓ and μ_s are the posterior means of the level and shape subspaces. Second, a hard baseline-anchoring constraint subtracts the shape decoder's output at z_s=$\mathbf{0}$, so the shape pathway contributes nothing at z_s=$\mathbf{0}$ and z_s is forced to encode deviations from the baseline trajectory rather than baseline level.

Engineered Feature Baseline. As an interpretable reference, we cluster hand-crafted trajectory features with the *same* unsupervised procedure applied to the deep latents (standardised features, GMM-full K=4), so the comparison is like-for-like. We report two variants. A *shape* baseline uses five trajectory-shape descriptors that carry no baseline-level information: early slope (OLS over days 0–3), late slope (days 7–10), curvature (late − early slope), P/F variability (standard deviation), and the number of velocity reversals. An *outcome-aware* baseline adds the two survival signals the deep models also receive through their auxiliary head, ICU death and length of stay (Section 5.2), which are legitimate features for this baseline, not the $\mathrm{TC}_{\mathrm{ref}}$ class label. For reference we additionally fit a *supervised* multinomial logistic classifier that predicts $\mathrm{TC}_{\mathrm{ref}}$ from the shape features on a held-out split; this is a signal-existence probe, not an unsupervised clustering baseline, and is reported separately as such. All variants use K=4 to match the clinical reference.

Training Details. All models are trained with the Adam optimiser (Kingma and Ba, 2015), gradient clipping (max norm = 5.0), and early stopping on validation loss, with linear β-warmup. Full hyperparameters (learning rates, batch sizes, hidden dimensions, dropout) are provided in Appendix D.

5.2. Survival Objective

Models B, D1–D3 include an auxiliary discrete competing-risk survival classifier, similar to Lee et al. (2018). At each patient's terminal event day, a softmax head predicts the event type (alive/censored, ICU death, or ICU discharge) from the latent embedding. The survival loss is a multi-class cross-entropy at the terminal event day only; patients who neither die

nor discharge contribute a censoring signal. This task encourages the latent representation to encode prognostically relevant information. Full details are in Appendix C.

5.3. Evaluation Metrics

Test-set embeddings are clustered using GMM-full (K=4, n_{init}=10) post hoc and matched to TC_{ref} via Hungarian assignment maximising macro-F1. We report: *Adjusted Rand Index* (ARI), *macro-F1*, *TC3 recall* (fraction of true TC3 in TC3-matched cluster), and *TC2 contamination* (fraction of TC2 in TC3-matched cluster; lower is better). Primary-benchmark metrics are reported as mean ± SD over 20 seeds (101–120) with paired test-set bootstrap 95% confidence intervals on the key gaps; the ancillary analyses that follow (external validation, disentanglement probes, reconstruction parity, and the high-certainty subset) retain the original three-seed runs (101–103). Model rankings are robust to clusterer choice (Spearman ρ=0.93 GMM-full vs. GMM-tied; ρ>0.80 across K=3–5). A Ridge regression severity probe (R^2 from each latent subspace to baseline P/F) quantifies disentanglement.

6. Results and Analysis

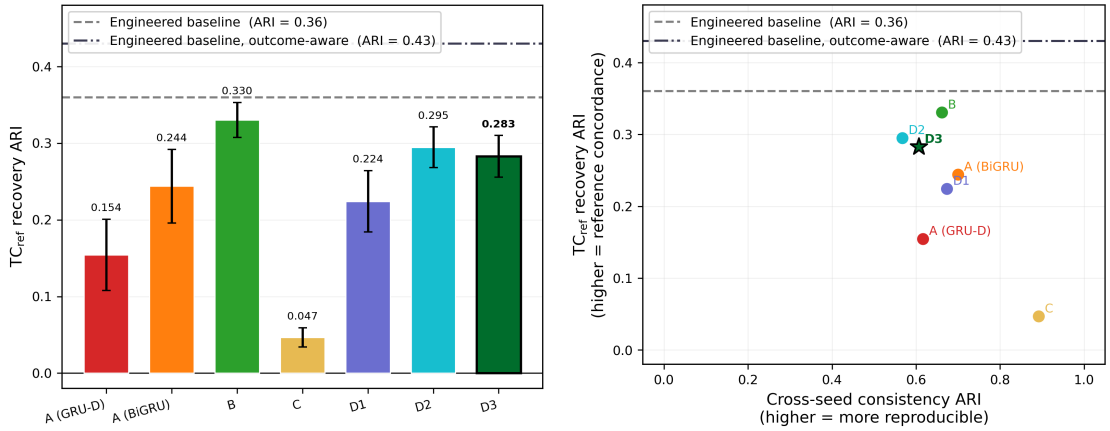
We now evaluate the same architectures on clinical AHRF data from MIMIC-IV, EUKC, and ENC. The synthetic experiment demonstration from Sec 3.2 predicted that VAE latent spaces organise along baseline severity and fail to separate trajectory classes with matched severity but divergent temporal shape. The results below confirm this and compare deep models with domain-engineered representations.

6.1. Overall Recovery of Reference Trajectory Classes TC_{ref}

We first assess the ability of each of the baseline models to recover the clinically validated reference trajectory classes TC_{ref} from (Marshall et al., 2026). Table 2 reports test-set performance (mean ± SD over 20 seeds, 101–120).

Deep and engineered representations recover the reference classes comparably. Overall recovery was modest (Table 2). Model B was the strongest deep model (ARI 0.330±0.023), but the shape-only engineered baseline was comparable (≈0.36), and the outcome-aware baseline, given the same survival information, performed better (0.43). A supervised classifier using the shape features achieved ARI 0.42 on held-out labels, confirming that trajectory information was available but not prioritised by unsupervised deep representations. Performance increased on the high-certainty TC_{ref} subset (posterior probability > 0.90; Appendix L), indicating that label uncertainty attenuated absolute ARI while largely preserving model ordering, although D2 overtook Model B on this subset. Survival-auxiliary models occupied the top three deep-model positions, while the BiGRU encoder outperformed GRU-D among reconstruction-only models (0.244 vs. 0.154).

Per-class performance and the TC2/TC3 litmus test. TC2 (stable persistence) is consistently the easiest class to recover, likely because it is the majority class (42.4%) with a relatively homogeneous flat trajectory. TC3 (biphasic) is the hardest: it requires capturing trajectory *shape* improvement through day 7 followed by deterioration. Model C achieves the highest TC3 recall (0.688±0.024) but the worst ARI (0.047), reflecting cluster collapse instead of genuine shape recovery.



A ARI by model (mean ± SD, 20 seeds). B Stability vs. validity (20 seeds).

Figure 2: TC$_{ref}$ recovery and stability across 20 seeds (GMM-full, K=4; MIMIC test set). **(A)** ARI by model; horizontal lines show the shape-only (0.36) and outcome-aware (0.43) engineered baselines. **(B)** Cross-seed clustering consistency versus TC$_{ref}$ ARI, demonstrating that stability is not a proxy for clinical validity.

Table 2: TC$_{ref}$ recovery on the MIMIC test set (GMM-full, K=4; mean ± SD over 20 seeds). TC3 Recall is the proportion of true TC3 assigned to its matched cluster; TC2 Contam. is the proportion of TC2 within that cluster (lower is better). Paired test-set bootstrap (B=2,000): the outcome-aware engineered baseline significantly exceeds the best deep model (95% CI on the ARI gap $[0.089, 0.173]$, excluding 0), whereas the 0.004 gap between Model B and the missingness-only baseline is not significant (95% CI includes 0); among the constrained models only D1's TC2-contamination reduction is robust.

Model	**ARI**	**Macro F1**	**TC3 Recall**	**TC2 Contam.↓**
B (SeqVAE+surv)	**0.330±0.023**	**0.547±0.034**	0.451±0.054	0.406±0.120
D2 (SSM+CR)	0.295±0.026	0.481±0.035	0.349±0.050	0.213±0.190
D3 (Anchored)	0.283±0.027	0.473±0.031	0.368±0.044	0.278±0.207
A (BiGRU)	0.244±0.048	0.457±0.019	0.434±0.024	0.488±0.061
D1 (Basis+CR)	0.224±0.040	0.421±0.028	0.431±0.059	**0.063±0.032**
A (GRU-D)	0.154±0.046	0.403±0.031	0.411±0.081	0.446±0.161
C (SplitLatent)	0.047±0.012	0.335±0.016	**0.688±0.024**	0.209±0.006

Internal selection metrics do not imply clinical validity. Model C was the most stable but least concordant with TC$_{ref}$ (Fig. 2B), and reconstruction loss was essentially unrelated to ARI (Pearson r=0.11; Appendix H). TC2/TC3 separability also peaked before the best-ELBO epoch, with an oracle shape-based stopping rule improving TC2/TC3 ARI by 0.089 over ELBO selection.

Nuisance baselines nearly match deep models. Missingness alone achieved ARI 0.326, only 0.004 below Model B, largely by identifying the shorter observation windows of TC1

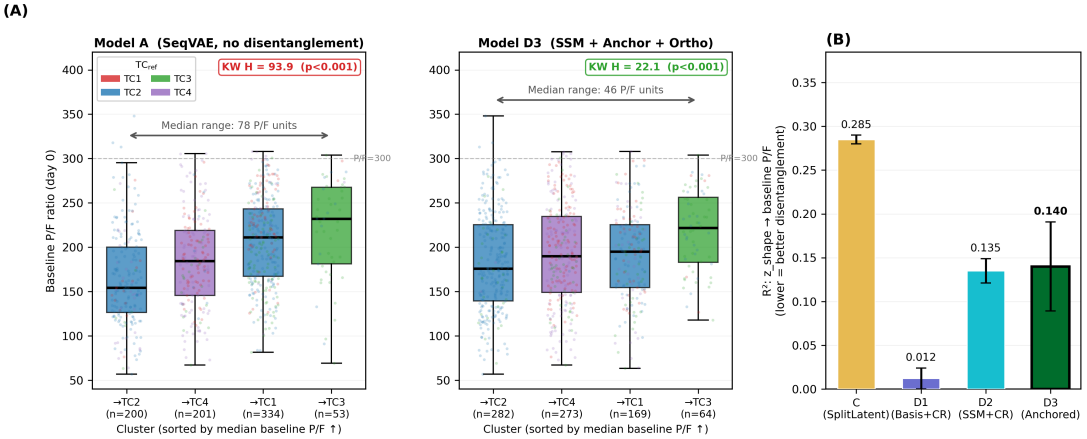


Figure 3: Severity diagnostics. **(A)** Day-0 P/F distributions across matched GMM-full (K=4) clusters for Models A and D3, coloured by dominant TC_{ref} class (which can differ from the →TC match label); Model A shows stronger baseline-severity stratification (Kruskal–Wallis H=93.9 versus 22.1). **(B)** Severity leakage into the shape subspace, $R^2(z_s)$: Model C 0.285, D1 0.012, and D3 0.140.

and TC4; it did not reliably distinguish TC3 from TC2. Overall ARI is therefore partly driven by the observation process, whereas TC2/TC3 separation remains the shape-sensitive test. No architecture was uniformly superior (Table 2): Model B achieved the highest overall ARI, while D1 produced the only robust reduction in TC2 contamination relative to B (0.063±0.032).

TC2 and TC3 have statistically indistinguishable baseline severity (day-0 P/F 187.9 vs. 186.9 mmHg, Welch p=0.85; separable at AUC 0.47, i.e. chance), so the severity axis that dominates the latent cannot tell them apart. Because the two classes are matched on the baseline-severity factor the representations actually encode, separating them requires trajectory *shape*, and shape is exactly what the representations miss: no deep model reliably recovers the distinction, and the synthetic experiment (Sec. 3.2) shows the failure is objective-driven. The frozen encoder achieves TC2/TC3 ARI = 0.024, no better than chance on this severity-matched pair.

6.2. Latent-Space Diagnostics

Figure 3A visualises the severity trap directly. Model A's clusters are strongly stratified by baseline P/F (H=93.9), mapping near-monotonically onto severity strata. Model D3's clusters show weaker stratification (H=22.1) with non-monotonic TC_{ref} ordering, indicating partitioning beyond the severity axis, though residual severity encoding persists ($R^2(z_s)$=0.140).

The severity probe (Fig. 3B; full results in Appendix E) reveals that Model C inverts the intended level/shape separation ($R^2(z_s)$=0.285, $R^2(z_\ell)$=0.022). D1 achieves near-complete disentanglement ($R^2(z_s)$=0.012) but lower ARI (0.224), demonstrating that *disentanglement is not sufficient*. This dissociation suggests that phenotype recovery requires both separating

severity from shape *and* retaining sufficient trajectory-shape information for downstream clustering; over-regularising the shape subspace can be as harmful as under-regularising it. Model B (ARI = 0.330) achieves phenotype recovery through the survival auxiliary instead of explicit disentanglement, despite substantial severity leakage ($R^2(z)$=0.438); the permutation control below suggests that the auxiliary improves latent geometry largely through regularisation rather than outcome-specific information.

Cluster trajectory profiles. Figure 12 (Appendix Q) shows cluster median P/F trajectories overlaid on $\mathrm{TC}_{\mathrm{ref}}$ medians. Model B's clusters show partial shape recovery but residual severity stratification. Model D3's clusters better track $\mathrm{TC}_{\mathrm{ref}}$ trajectory shapes, especially the biphasic TC3 pattern, though its overall ARI (0.283) is lower than B's (0.330), illustrating the trade-off between unconstrained use of all signal (B) and explicit disentanglement (D3).

Shape is present in the latents but not recovered by clustering. The severity trap is a failure of *geometry*, not of representational capacity: the information needed to separate the hardest class pair is present in every latent, but it is never the dominant axis that unsupervised clustering selects. A supervised linear classifier (5-fold CV logistic regression) trained on each frozen latent to separate TC2 from TC3 succeeds at essentially the raw-data level: AUC 0.905–0.954 across all seven models, well above the day-0 P/F floor of 0.469 (chance) and matching or exceeding the raw-trajectory reference of 0.918. Yet unsupervised GMM clustering of the same latents recovers TC2/TC3 poorly (TC2-contamination 0.21–0.54). Trajectory shape is thus encoded but sub-dominant, and clustering follows the severity axis instead.

The survival auxiliary acts as a regulariser, not an outcome channel. Model B's competing-risk auxiliary is the only objective that lifts recovery above the reconstruction-only models, raising the concern that it merely aligns the latent with the outcome-derived reference. A permutation control argues otherwise: in a three-seed control, retraining Model B with patient event labels randomly permuted (destroying real prognostic signal while preserving the auxiliary's regularising presence) leaves most of its advantage intact (ARI 0.344→0.279; it does not fall to the reconstruction-only range of 0.15–0.24), and a logistic-regression probe on the placebo head's predicted event-risk and cumulative-incidence outputs still separates TC2/TC3 (AUC 0.92, versus 0.80 for the same probe on the real head). The auxiliary's benefit is therefore largely a regularisation effect, and the TC2/TC3 signal it appears to carry is inherited from the latent geometry rather than learned from outcomes.

Recovery is not an artefact of forcing K=4. The K=4 constraint does not create the failure. Across a full K=2–8 sweep no representation exceeds ARI ≈0.35 at any K (best 0.349, Model B at K=3), so the ceiling is K-independent; and clustering each latent at its own BIC-optimal $K^\star$ (Table 16, Appendix Q) does not improve recovery (ARI 0.04–0.27), instead partitioning by baseline-severity *level* (η^2=0.19–0.49). The unconstrained Model A (GRU-D) collapses to $K^\star$=2, and that split *is* the severity axis (level separates the two clusters at AUC 0.96). The models' own preferred geometry is the severity axis; forcing K=4 merely makes it legible against a clinical reference.

6.3. External Validation across Multiple Cohorts

We evaluate whether trajectory representations learned on MIMIC-IV transfer to external cohorts. To test generalisation, we apply frozen MIMIC-trained encoders to the EUKC and ENC cohorts and refit the GMM-full clustering model ($K = 4$) on each target dataset. This allows us to assess both phenotype recovery across cohorts and whether disentanglement between severity and trajectory shape transfers under distribution shift. Across the three best-performing models, Model B achieves the highest external ARI on EUKC (0.250), and on ENC, B, D2, and D3 perform comparably (0.298-0.302; full results for all seven models in Appendix K).

Reference class recovery partially generalises across cohorts, but disentanglement degrades under distribution shift. For the anchored model D3, recovery of the reference trajectory phenotypes remains comparable between MIMIC and ENC (MIMIC ARI = 0.298 vs. ENC ARI = 0.302) but drops substantially on the EUKC cohort (ARI = 0.170; Fig. 13, Appendix Q). At the same time, severity leakage into the trajectory-shape subspace increases under distribution shift: $R^2(z_s)$ rises from 0.140 on MIMIC to 0.272 on ENC and 0.348 on EUKC (Table 17, Appendix Q). These results indicate that disentanglement learned during training is only partially robust to dataset shift. In particular, the increase in $R^2(z_s)$ suggests that baseline severity re-enters the trajectory subspace in external cohorts, weakening the separation between severity and trajectory shape. This finding implies that domain adaptation or site-aware constraints may be necessary for trajectory representations to generalise reliably across clinical settings.

External results are not an artefact of porting the reference. Re-scoring the frozen encoders against natively-fit CRLCMMs at each site (rather than the ported MIMIC reference) leaves the external picture essentially unchanged, so the conclusions are not a circularity artefact (Appendix K).

6.4. Beyond VAEs: the severity trap across sequential architectures

Our primary object of study is the sequential VAE family, but the severity trap is a property of the training *objective* rather than of variational autoencoding specifically. To show this, we evaluate three further families of sequential representation learner on the same MIMIC-IV data and evaluation pipeline: (i) a *masked-reconstruction transformer* (a Transformer encoder trained to predict a held-out fraction of observed days); (ii) a *diffusion autoencoder*, probing its semantic latent; and (iii) a *contrastive* encoder (SimCLR-style instance discrimination on augmented trajectories), evaluated both with standard, level-preserving augmentations and with an explicit *baseline-shift* augmentation that adds a random constant to the whole series, forcing invariance to baseline level. All encoders match the VAEs' backbone capacity, latent dimension, and post-hoc clustering; details are given in Appendix P.

Table 3 reports the severity probe (R^2 from the embedding to day-0 P/F) and $\mathrm{TC}_{\mathrm{ref}}$ recovery. Every *reconstruction-based* objective falls into the trap: the masked transformer encodes severity almost identically to the VAEs (R^2=0.40), and the diffusion autoencoder shows the strongest dominance of any architecture (R^2≈0.998). Contrary to the intuition that contrastive objectives avoid baseline magnitude by construction, the standard contrastive encoder encodes day-0 severity *almost perfectly* (R^2=0.973, higher than any VAE): its

Table 3: The severity trap across sequential architectures (MIMIC-IV test split; contrastive, transformer and diffusion over 3 seeds). Severity R^2: Ridge probe from the embedding to day-0 P/F (higher = more severity-dominated). TC_{ref} ARI is the full four-class recovery; TC2/TC3 ARI is on the TC2∪TC3 subset. Rows ordered by severity encoding. Every reconstruction- or level-preserving objective encodes severity strongly; only the level-invariant contrastive objective sheds it, and only partially.

Encoder / objective	$R^2(z \to$ d0 P/F$)$	TC_{ref} **ARI**	**TC2/TC3 ARI**
Diffusion autoencoder	0.998	—	0.08–0.38
Contrastive, level-preserving (no shift)	0.973	0.257	0.188
Contrastive, level-invariant (baseline-shift)	0.468	0.341	0.226
Sequential VAE (Model B, primary)	0.438	0.330	—
Masked-reconstruction transformer	0.398	0.302	0.273
Engineered (outcome-aware)	—	*0.43*	—

augmentations preserve baseline level, so instance discrimination is most easily solved by reading off severity. The trap is escapable only by explicitly removing the nuisance axis: a baseline-shift augmentation spanning the severity range drops severity encoding from 0.97 to 0.47 (and to 0.37 under a wider shift), yet the escape is only *partial*: TC2/TC3 recovery rises only to ARI ≈0.21–0.26, and the best of these architectures reaches four-class TC_{ref} ARI 0.341, still below the outcome-aware engineered baseline (0.43). Invariance to baseline level must therefore be built in deliberately rather than expected to emerge.

7. Discussion

In this multi-cohort benchmark of trajectory representation learning in AHRF, commonly used sequential VAE architectures learned latent spaces dominated by baseline oxygenation severity rather than trajectory shape. Auxiliary competing-risk supervision improved agreement with the externally validated reference trajectories but did not eliminate this tendency, and an interpretable engineered-feature baseline clustered the same unsupervised way remained competitive, exceeding the best deep model once given the same survival signal (ARI 0.43 vs. 0.330). The relevant trajectory information is present in the data but was not prioritised by reconstruction-focused objectives when baseline severity dominates input variance.

Severity-dominant latent organisation and its partial remedies. We use the term *severity trap* to describe a setting in which ELBO-trained sequential VAE latent spaces organise along baseline severity (∼50% of P/F variance) rather than temporal trajectory shape. This differs from posterior collapse because the latents remain informative but allocate their dominant geometry to severity rather than trajectory shape. The controlled synthetic experiment (Section 3.2) localises the failure to the ELBO objective itself: a model trained from scratch on synthetic data exhibits the same TC2/TC3 discrimination failure (ARI = 0.047). While the VAE family is our primary object of study, the trap is not confined to it: a masked-reconstruction transformer and a diffusion autoencoder fall into it as severely, and a standard contrastive encoder escapes only partially, and only when

made explicitly invariant to baseline level (Section 6.4). The failure therefore tracks the training objective (reconstruction- and level-preserving objectives fall in) rather than deep representation learning *per se*.
Competing-risk supervision partially improved recovery, although some benefit may reflect alignment with a reference model that also incorporates competing risks. The failure to recover the severity-matched TC2 and TC3 classes nevertheless shows that outcome supervision did not reliably recover trajectory shape.

Cross-seed clustering stability (up to 0.89) and reconstruction loss both fail to predict TC_{ref} recovery, and a missingness-only baseline (ARI 0.326) nearly matches the best deep model, exposing the observation process as a second nuisance axis alongside severity (Section 6). Unsupervised trajectory phenotyping models should therefore be validated against external clinical references with nuisance-aware baselines, not selected on internal generative metrics.

Clinical relevance. The reference trajectory model suggests that early class probabilities may discriminate later oxygenation course better than day-3 P/F ratio alone: day-3 trajectory-class probabilities predict 14-day mortality at AUC 0.74 versus 0.55 for the P/F ratio alone (Marshall et al., 2026), supporting early identification of patients whose condition will deteriorate before clinical signs are apparent. Trajectory class probabilities at day 3 could serve as a dynamic stratification variable for clinical trials, capturing predicted clinical course instead of static baseline severity. The severity trap, however, has a measurable cost for exactly this use: in a prognostic-enrichment simulation, deep-model clusters are no better than engineered features and need roughly 3–3.5× the per-arm sample size of the oracle reference to power a deterioration trial (Appendix O). The severity trap causes deep architectures to stratify patients by baseline severity rather than clinical evolution, precisely the static classification trajectory phenotyping aims to supersede. This has a clinical analogue in *anchoring bias* (fixating on a patient's initial presentation and under-weighting their subsequent course), which makes the failure mode intuitive at the bedside.

Limitations. TC_{ref} is a reference standard derived from a specific modelling framework, not ground truth; a sensitivity analysis on the high-certainty subset (posterior probability >0.90) yields substantially higher ARIs with largely preserved model ordering (Appendix L), evidence that reference-label noise attenuates absolute performance but does not distort model comparisons. Prospective clinician adjudication of the derived archetypes would further strengthen the reference labels and is a valuable next step. The primary analysis uses univariate P/F trajectories, but the trap arises from the interaction between the ELBO objective and the data's variance structure rather than input dimensionality: preliminary multivariate ablations (Appendix M) show that adding severity-correlated variables deepens rather than resolves it. External validation shows that disentanglement degrades under distribution shift, with severity re-entering the shape subspace at external sites; domain adaptation strategies for trajectory representations remain to be developed. The synthetic dataset is deliberately simple; whether the trap persists in more complex synthetic settings with realistic covariate structure remains to be tested.

Future directions. Promising solutions include level-invariant contrastive losses, changepoint-aware encoders for non-monotonic trajectories, and discrete latent models that force commitment to trajectory archetypes. Because missingness is itself predictive and disentanglement

degrades across sites, future work should also model the observation process and incorporate domain adaptation.

8. Conclusion

Reconstruction- and level-preserving sequential representation learners consistently prioritised baseline oxygenation severity over trajectory shape across three AHRF cohorts. We term this failure mode the *severity trap*: it was not detected by reconstruction error or clustering stability, and nuisance-aware engineered baselines matched or exceeded the deep models. Trajectory phenotyping therefore requires clinically grounded external validation and objectives that explicitly suppress baseline-level and observation-process shortcuts.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. D.C.M. is supported by an MRC Clinical Research Training Fellowship (award MR/Y000404/1) and the Mittal Fund at Cleveland Clinic Philanthropy.

References

Alexander Alemi, Ben Poole, Ian Fischer, et al. Fixing a broken ELBO. In *Proceedings of ICML*, 2018.

Lieuwe D. J. Bos, Michael Sjoding, Pratik Sinha, Sivasubramanium V. Bhavani, Patrick G. Lyons, Alice F. Bewley, Michela Botta, Anissa M. Tsonas, Ary Serpa Neto, Marcus J. Schultz, Robert P. Dickson, and Frederique Paulus. Longitudinal respiratory subphenotypes in patients with COVID-19-related acute respiratory distress syndrome: results from three observational cohorts. *The Lancet Respiratory Medicine*, 9(12):1377–1386, 2021. doi: 10.1016/S2213-2600(21)00365-9.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, et al. Generating sentences from a continuous space. In *Proceedings of CoNLL*, pages 10–21, 2016.

Carolyn S. Calfee, Kevin Delucchi, Polly E. Parsons, B. Taylor Thompson, Lorraine B. Ware, Michael A. Matthay, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *The Lancet Respiratory Medicine*, 2(8):611–620, 2014. doi: 10.1016/S2213-2600(14)70097-9.

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

Hui Chen, Qian Yu, Jianfeng Xie, Songqiao Liu, Chun Pan, Ling Liu, et al. Longitudinal phenotypes in patients with acute respiratory distress syndrome: a multi-database study. *Critical Care*, 26:340, 2022. doi: 10.1186/s13054-022-04211-w.

Xi Chen, Diederik P. Kingma, Tim Salimans, et al. Variational lossy autoencoder. In *Proceedings of ICLR*, 2017.

Katie R. Famous, Kevin Delucchi, Lorraine B. Ware, et al. Acute respiratory distress syndrome subphenotypes respond differently to randomized fluid management strategy. *American Journal of Respiratory and Critical Care Medicine*, 195(3):331–338, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV (version 1.0), 2021. RRID:SCR_007345.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Changhee Lee, William R. 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*, 2018. doi: 10.1609/aaai.v32i1.11842.

Xiaofei Li, Lei Gao, Lian Li, et al. Identifying vital sign trajectories to predict 28-day mortality of critically ill elderly patients with ARDS. *Respiratory Research*, 25:10, 2024.

Ramanathan R. Ling, Mallikarjuna Ponnapa Reddy, Ashwin Subramaniam, et al. Epidemiology of acute hypoxaemic respiratory failure in Australian and New Zealand intensive care units during 2005–2022: a binational, registry-based study. *Intensive Care Medicine*, 50 (11):1861–1872, 2024. doi: 10.1007/s00134-024-07609-y.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

Corbin Loewen, Brenden Dufault, Owen Mooney, Kendiss Olafson, and Duane J. Funk. Identification of four latent classes of acute respiratory distress syndrome using PaO_2/FiO_2 ratio: an observational cohort study. *Scientific Reports*, 14:2042, 2024. doi: 10.1038/s41598-024-52243-9.

James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Don't blame the ELBO! A linear VAE perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, 2019.

Dominic C Marshall, Ashleigh D Green, Matthieu Komorowski, Brijesh V Patel, Sonali Parbhoo, and David B Antcliffe. Reproducible clinical archetypes in acute respiratory failure: a multi-cohort trajectory analysis. *Intensive Care Medicine*, pages 1–12, 2026.

Tài Pham, Antonio Pesenti, Giacomo Bellani, et al. Outcome of acute hypoxaemic respiratory failure: insights from the LUNG SAFE study. *European Respiratory Journal*, 57(6):2003317, 2021. doi: 10.1183/13993003.03317-2020.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, 2019.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

Pratik Sinha, Kevin L. Delucchi, B. Taylor Thompson, Daniel F. McAuley, Michael A. Matthay, Carolyn S. Calfee, et al. Latent class analysis of ARDS subphenotypes: a secondary analysis of the Statins for Acutely Injured Lungs from Sepsis (SAILS) study. *Intensive Care Medicine*, 44(11):1859–1869, 2018. doi: 10.1007/s00134-018-5378-3.

Wei Song, Guoqiang Cai, Yongchun Zhou, et al. Trajectory patterns of mechanical power and prognosis in ARDS. *European Journal of Medical Research*, 30:39, 2025.

Patrick J. Thoral, Jan M. Peppink, Ronald H. Driessen, Eric J. G. Sijbrands, Erwin J. O. Kompanje, Lewis Kaplan, et al. Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine joint data science collaboration: the Amsterdam University Medical Centers Database (AmsterdamUMCdb) example. *Critical Care Medicine*, 49(6):e563–e577, 2021. doi: 10.1097/CCM.0000000000004916.

Lei Wang, Wei Zhang, Yubiao Ge, et al. Risk stratification of patients with ARDS complicated with sepsis using lactate trajectories. *BMC Pulmonary Medicine*, 22:347, 2022.

Appendix A. Theoretical Insights On the Severity Trap in Deep Trajectory Phenotyping

So far, we have shown empirically that deep trajectory representations often organise along baseline severity rather than trajectory shape. We provide a rate–distortion analysis of the ELBO objective under a simplified trajectory factor model. Specifically, we analyse the ELBO under a simplified generative model of clinical trajectories and show that, when baseline severity contributes more reconstruction-relevant variance than trajectory shape, the optimal representation preferentially encodes severity. This provides a mechanistic explanation for the severity trap defined in Definition 1.

Trajectory factor model. We assume that each trajectory can be decomposed into a baseline severity component and a temporal trajectory component,

$$x_t = s + f_t(\tau) + \varepsilon_t, \qquad t = 1, \dots, T,$$

where s denotes baseline severity, τ indexes trajectory shape, $f_t(\tau)$ describes the temporal dynamics, and $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is observation noise. The variables s, τ, and ε_t are assumed independent. Let

$$\sigma^2_{\text{shape}} = \sum_{t=1}^{T} \text{Var}(f_t(\tau))$$

denote the total variance attributable to trajectory shape.

ELBO as a rate–distortion objective. The VAE objective maximises the evidence lower bound

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \beta \text{KL}(q_\phi(z \mid x) \| p(z)).$$

Under a Gaussian decoder, this objective is equivalent to minimising a reconstruction distortion term together with a rate penalty induced by the KL divergence. The KL term limits the information that the latent representation can encode about the input, forcing the model to allocate representational capacity to those factors of variation that yield the largest reduction in reconstruction error.

Proposition 2 (Severity dominance under ELBO) *Consider the trajectory model*

$$x_t = s + f_t(\tau) + \varepsilon_t, \qquad t = 1, \ldots, T.$$

Assume s, τ, and ε_t are independent, $s \sim \mathcal{N}(0, \sigma_s^2)$, and $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. Let z denote the latent representation learned under the ELBO objective with limited effective capacity induced by the KL penalty. If

$$T\sigma_s^2 > \sum_{t=1}^{T} \text{Var}(f_t(\tau)),$$

then encoding baseline severity yields a larger reduction in expected reconstruction error than encoding trajectory shape. Consequently, the ELBO objective preferentially allocates representational capacity to severity.

Proof sketch. If baseline severity is not encoded, it contributes an expected squared reconstruction error of order $T\sigma_s^2$, since the same offset appears at every time point. Encoding s removes this error. By contrast, the maximal reduction obtainable from encoding trajectory shape is bounded by

$$\sum_{t=1}^{T} \text{Var}(f_t(\tau)).$$

Because the KL term penalises all encoded information equally, the most efficient use of limited representational capacity is to encode the factor that yields the largest reduction in reconstruction error. Thus, when $T\sigma_s^2$ exceeds the variance attributable to trajectory shape, the ELBO objective favors encoding severity before shape.

Corollary 3 (Failure to separate shape-defined classes) *Suppose two trajectory classes share the same distribution of baseline severity but differ in trajectory shape. If a learned representation satisfies*

$$I(z; s) \gg I(z; \tau),$$

then the class-conditional latent distributions $p(z \mid C_1)$ and $p(z \mid C_2)$ overlap substantially, and clustering based on z cannot reliably separate the classes.

Appendix B. Synthetic Data Generation Details

Each synthetic class is defined by a baseline mean and a shape function. TC1-like: μ=225 mmHg, exponential improvement ($\Delta\text{PF}(t) = 55(1-e^{-t/3.5})$). TC2-like: μ=175 mmHg, flat (no temporal change). TC3-like: μ=175 mmHg, biphasic (+60 mmHg peak at day 7, returning to baseline by day 13). TC4-like: μ=155 mmHg, linear decline ($\Delta\text{PF}(t) = -9t$). Gaussian noise (σ=9 mmHg) and ∼30% random missingness were added to all classes.

Method	**ARI (all)**	**ARI (TC2/TC3)**
Model B (frozen) + GMM	0.160	0.024
Model A (from scratch) + GMM	0.194	0.047

Table 4: Synthetic severity trap. Both VAE models fail to separate TC2/TC3 (ARI ≈ 0) even though the two classes differ only in trajectory shape.

Appendix C. Survival Objective Details

For each patient i with D_i observed days, let T_i denote the event day and $E_i \in \{0, 1, 2\}$ the event type (alive/censored, ICU death, ICU discharge). The survival head outputs $\hat{p}(E{=}e \mid z, d) = \text{softmax}_e(\mathbf{W}_e z + b_e)$, trained with cross-entropy at the terminal event day: $\mathcal{L}_{\text{surv}} = -\log \hat{p}(E{=}E_i \mid z_i, T_i)$.

Appendix D. Training Hyperparameters

Model	**Encoder**	**Decoder**	**Latent**	**Survival**	**Orthog.**
A (GRU-D)	GRU-D	GRU-AR	$z = 32$	–	–
A (BiGRU)	BiGRU	GRU-AR	$z = 32$	–	–
B	GRU-D	GRU-AR	$z = 16$	✓	–
C	GRU-D	GRU-AR	$z_\ell + z_s = 8 + 8$	–	–
D1	GRU-D	Basis ($n = 5$)	$z_\ell + z_s = 8 + 8$	✓	–
D2	GRU-D	SSM	$z_\ell + z_s = 8 + 8$	✓	–
D3	GRU-D	SSM	$z_\ell + z_s = 8 + 8$	✓	Cross-cov + anchor

Table 5: Architectures evaluated in the trajectory representation benchmark. Models differ along three design axes: latent structure (shared vs. level/shape split), auxiliary survival objective, and orthogonality regularisation between latent subspaces. Full training hyperparameters are listed below

Models A–C: 2-layer encoder, 128 hidden units, dropout 0.15, lr 10^{-3}, batch 128, β=0.5 (A) / 0.1 (B) / 0.3 (C), 20-epoch warmup, up to 200 epochs. D1: same encoder, 50-epoch minimum, basis smoothness regularisation. D2/D3: 96 hidden units, dropout 0.10, lr $4{\times}10^{-3}$, batch 1024, β=0.1, 15-epoch warmup, d_s=8. D2 applies no orthogonality penalty. D3: $\lambda_\perp$=0.1 (cross-covariance orthogonality penalty); baseline anchoring is a hard architectural constraint with no weight. Weight decay 10^{-2} for D2/D3.

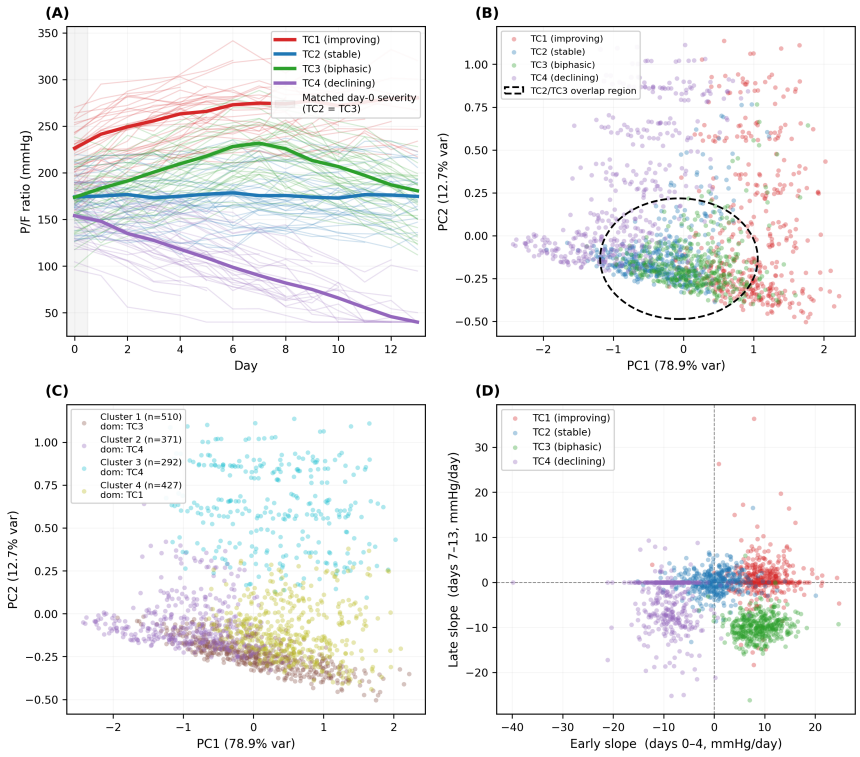


Figure 4: Synthetic demonstration of the severity trap. **(A)** Four synthetic trajectory classes; TC2 and TC3 share an identical day-0 P/F distribution (median 175 mmHg, SD 28), so only shape distinguishes them. **(B)** The frozen Model B latent space, coloured by true class, organises along a single severity axis (PC1 = 78.9% of variance); TC2 and TC3 are spatially interleaved (dashed overlap region). **(C)** GMM clusters of the same latent partition by severity level and place TC2/TC3 together. **(D)** Engineered early- versus late-slope features separate the classes by construction in this shape-only cohort.

Appendix E. Severity Disentanglement Probe

Table 6: R^2 of Ridge regression from each latent subspace to baseline P/F (mean ± SD, 3 seeds). Lower $R^2(z_s)$: better disentanglement.

Model	$R^2(z)$	$R^2(z_\ell)$	$R^2(z_s)$ ↓
D1 (Basis+CR)	0.230±0.004	0.217±0.003	**0.012±0.012**
D2 (SSM+CR)	0.440±0.010	0.266±0.021	0.135±0.014
D3 (Anchored)	0.441±0.007	0.258±0.011	0.140±0.051
C (SplitLatent)	0.313±0.008	0.022±0.004	0.285±0.005
B (SeqVAE+surv)	0.438±0.023	—	—
A (GRU-D)	0.273±0.016	—	—

Appendix F. K-Selection Justification

F.1. Raw-Data Evidence for K=4

Raw-trajectory analysis (DPGMM + BIC elbow) supports K=4 as the parsimonious archetype count: the non-parametric DPGMM converges to four effective components across all α values tested, and the BIC shows its largest single-step drop at K=3→4 (Figure 5).

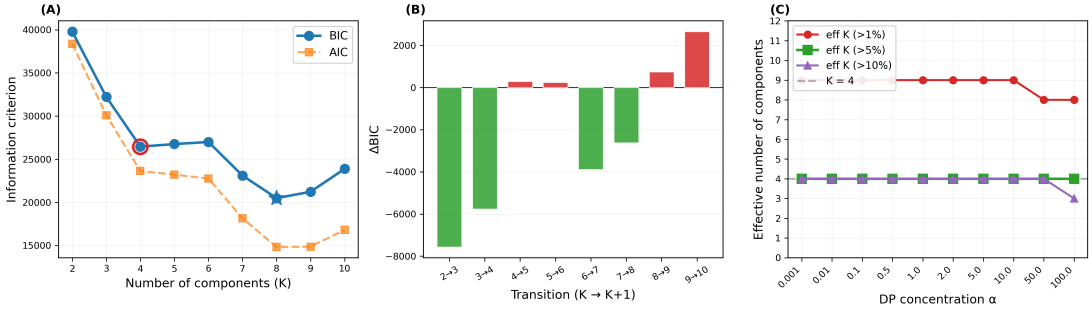


Figure 5: BIC elbow and DPGMM effective K on raw P/F trajectories (MIMIC-IV). DPGMM converges to K=4 across five orders of magnitude of α.

However, the data support 4–6 plausible granularities. The BIC global minimum on MIMIC lies at K=8 (vs. K=4 on the smaller EUKC cohort), reflecting that MIMIC's larger sample size resolves sub-clusters within archetypes particularly within TC2 (stable persistence), which contains heterogeneous ventilation-weaning profiles that could reasonably be split further. We adopt K=4 because: (i) it matches the CRLCMM reference and allows direct comparison with the clinical gold standard; (ii) it is the most reproducible granularity (ranking correlations ρ>0.80 up to K=5, declining above K=6); (iii) at K>4, the new clusters correspond to splits of TC2, not to novel archetypes.

F.2. Latent-Space K-Selection

Figure 6 shows BIC and ICL curves in the latent space for each model, and Figure 7 shows the DPMM-selected effective K.

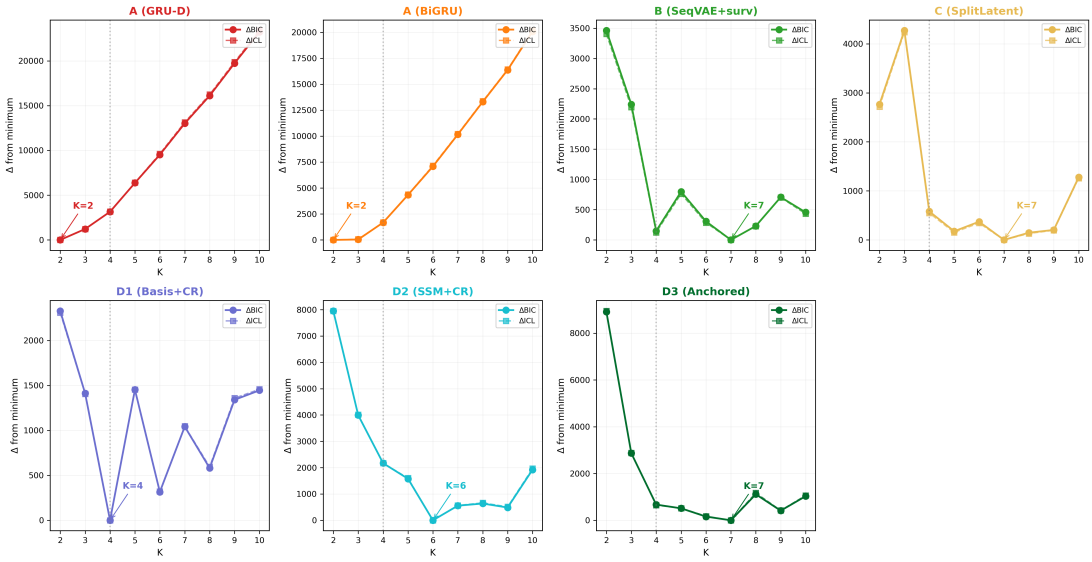


Figure 6: BIC (solid) and ICL (dashed) in latent space, per model, $K \in \{2, \ldots, 8\}$.

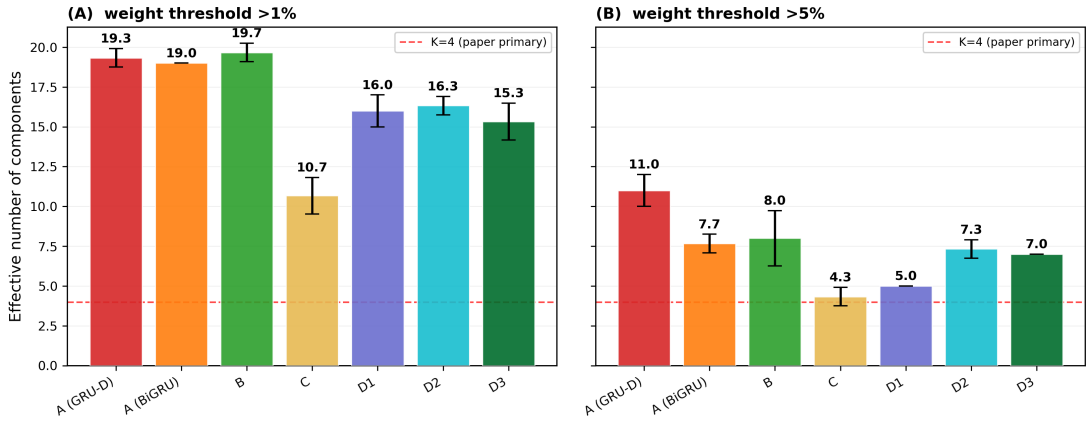


Figure 7: DPMM effective K on model latent spaces.

- D1 is the only model where the latent-space BIC minimum falls at K=4.
- D2/D3 show clear K=4 elbows in latent-space BIC.
- Model A (GRU-D) prefers K=2 (A (BiGRU) prefers K=3): the latent space collapses toward a dominant severity axis that only supports two well-separated clusters.
- Survival-auxiliary models (B, D1, D2, D3) support $K \geq 4$ structure in the latent space, consistent with their better $\mathrm{TC}_{\mathrm{ref}}$ recovery.

Appendix G. CRLCMM Reference Classes and Posterior Certainty

The reference trajectory classes $\mathrm{TC}_{\mathrm{ref}}$ are taken from the competing-risk latent class mixed model (CRLCMM) of Marshall et al. (2026); we restate its specification here for completeness. Let y_{it} be the P/F ratio for ICU stay i on day $t \in \{0, \ldots, 13\}$ and $c_i \in \{1, \ldots, G\}$ its latent trajectory class. The model jointly specifies three components.

Longitudinal submodel. Conditional on class $c_i{=}g$, the P/F trajectory follows a class-specific quadratic mixed model with a subject-specific random intercept and slope,

$$y_{it} \mid (c_i{=}g) = \beta_{0g} + \beta_{1g}\, t + \beta_{2g}\, t^2 + u_{0i} + u_{1i}\, t + \varepsilon_{it}, \tag{3}$$

with $(u_{0i}, u_{1i})^\top \sim \mathcal{N}(\mathbf{0}, \Sigma_u)$ and $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2)$. The class-specific fixed effects $(\beta_{0g}, \beta_{1g}, \beta_{2g})$ capture the four trajectory shapes; the random intercept and slope absorb within-patient correlation.

Competing-risks submodel. ICU death ($k{=}1$) and ICU discharge ($k{=}2$) are treated as competing events with class-specific, cause-specific Weibull baseline hazards,

$$\lambda_i^{(k)}(t \mid c_i{=}g) = \lambda_{0g}^{(k)}(t), \qquad k \in \{1, 2\}, \tag{4}$$

which share the subject's latent class with the longitudinal submodel, so that truncation of the trajectory by death or discharge is modelled rather than assumed missing-at-random, the key departure from a standard latent class mixed model.

Class membership and assignment. Class priors follow a multinomial logit, $\pi_g = \Pr(c_i{=}g) = e^{\xi_g} / \sum_{g'} e^{\xi_{g'}}$, and all parameters are estimated jointly by maximum likelihood via the Expectation–Maximisation algorithm (`lcmm::Jointlcmm`). Each stay's posterior is proportional to the prior times the longitudinal and survival likelihoods,

$$\Pr(c_i{=}g \mid \mathbf{y}_i, T_i, \delta_i) \;\propto\; \pi_g\, f(\mathbf{y}_i \mid c_i{=}g)\, \left[\lambda_{ig}^{(\delta_i)}(T_i)\right]^{\mathbf{1}[\delta_i>0]} S_{ig}(T_i), \tag{5}$$

where T_i is the event/censoring day, $\delta_i \in \{0, 1, 2\}$ the event type, and S_{ig} the class-specific overall survival. The hard reference label is the maximum-a-posteriori class $\hat{c}_i = \arg\max_g \Pr(c_i{=}g \mid \cdot)$, which is the benchmark target $\mathrm{TC}_{\mathrm{ref}}$.

Class enumeration. G was selected by BIC, relative entropy, the integrated completed likelihood, a minimum class size $> 10\%$, and bootstrap stability; $G{=}4$ satisfied all criteria, yielding TC1 (early recovery), TC2 (stable persistence), TC3 (biphasic), and TC4 (rapid decline). The four classes are corroborated across paradigms: de-novo four-class fits reproduce them in the external cohorts, and a model-free consensus clustering of the raw MIMIC trajectories independently supports four classes with matching shapes (cross-method ARI 0.27–0.36, as expected for distinct models). In the lower-mortality ENC cohort BIC favours $G{=}5$; $G{=}4$ is retained throughout for comparability.

Figure 8 summarises the $\mathrm{TC}_{\mathrm{ref}}$ trajectory profiles, class proportions across cohorts, and assignment certainty.

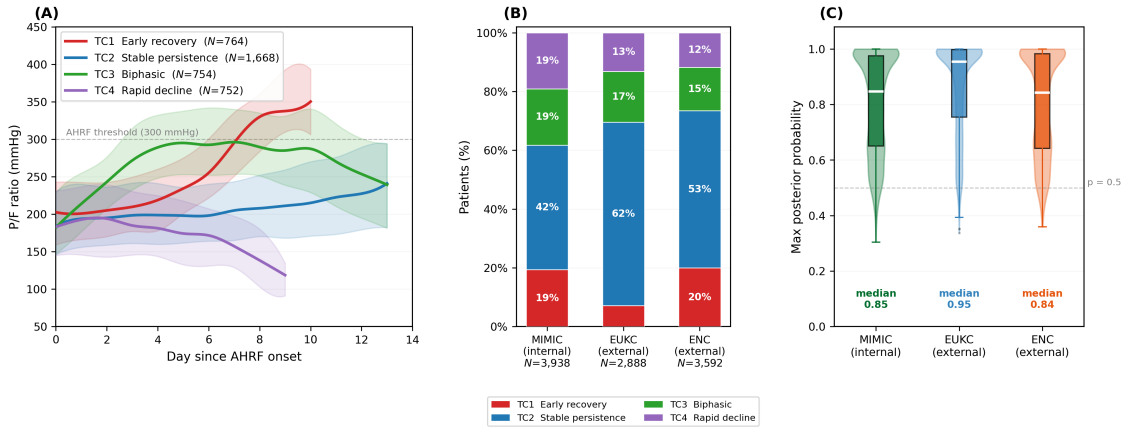


Figure 8: **(A)** MIMIC-IV TC_{ref} median P/F trajectories with IQR shading (days 0–13). AHRF threshold (300 mmHg) shown dashed. **(B)** Class membership proportions across cohorts; EUKC shows a higher proportion of TC2 (stable persistence), consistent with its ICU care-pathway characteristics. **(C)** CRLCMM assignment certainty (max posterior probability per patient). Median certainty: MIMIC 0.847, EUKC 0.954, ENC 0.843, confirming that the MIMIC-derived CRLCMM transfers with high confidence to both external cohorts.

Appendix H. Reconstruction Parity

Figure 9 and Table 7 show that reconstruction quality does not predict phenotype recovery.

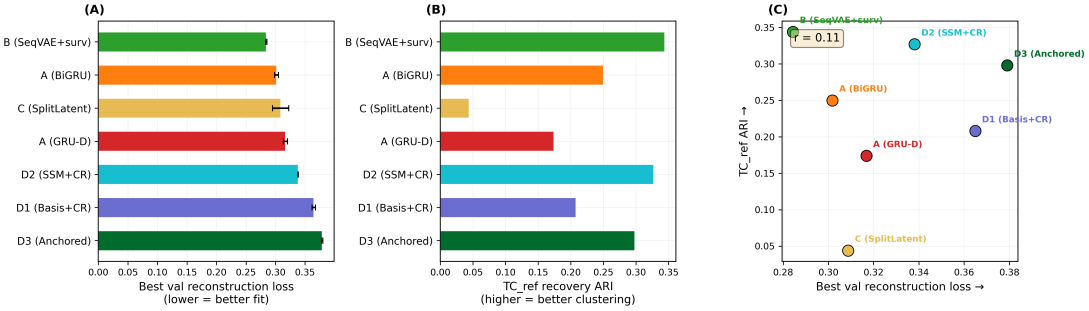


Figure 9: Reconstruction quality (best validation masked MSE) vs. TC_{ref} ARI. Pearson r=0.11, reinforcing that reconstruction fidelity does not predict cluster validity.

Table 7: Reconstruction parity. Best validation reconstruction loss (masked MSE, lower is better ↓) and corresponding test-set ARI. Mean ± SD over three seeds; Model B's ARI here (0.344) is the 3-seed value, whereas the 20-seed primary benchmark reports 0.330 (Table 2). Pearson $r = 0.11$ between reconstruction quality and ARI, confirming that cluster validity is not driven by reconstruction fidelity.

Model	Best Val. Recon. ↓	ARI
B (SeqVAE+surv)	**0.284 ± 0.001**	**0.344**
A (BiGRU)	0.302 ± 0.003	0.250
C (SplitLatent)	0.309 ± 0.014	0.044
A (GRU-D)	0.317 ± 0.004	0.174
D2 (SSM+CR)	0.338 ± 0.000	0.327
D1 (Basis+CR)	0.365 ± 0.003	0.208
D3 (Anchored)	0.379 ± 0.001	0.298

Appendix I. Variance Decomposition

Figure 10 and Table 8 decompose total cohort P/F trajectory variance into orthogonal polynomial components using a cross-patient decomposition.

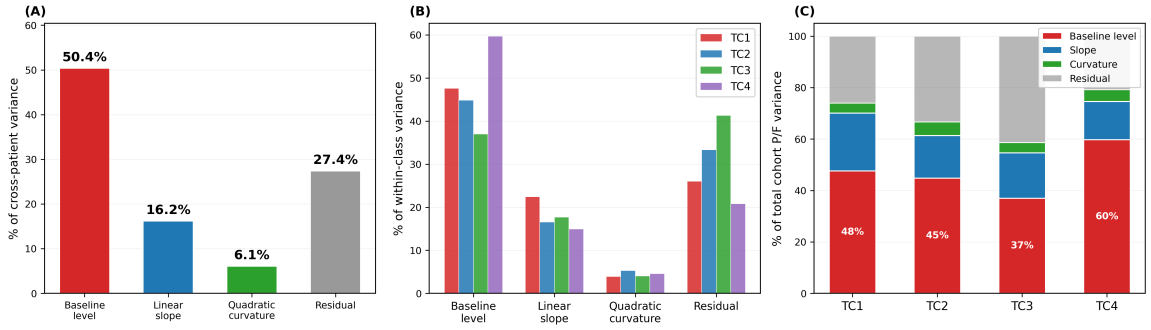


Figure 10: Cross-patient P/F trajectory variance decomposition by orthogonal polynomial basis (intercept, slope, curvature). **(A)** Overall decomposition across all patients: baseline severity accounts for 50.4% of total cohort variance. **(B)** Decomposition stratified by TC_{ref} class. **(C)** Stacked breakdown of total variance by TC class, showing that TC3 has the lowest baseline severity component (37%), meaning a greater proportion of its variance arises from trajectory dynamics consistent with its biphasic shape.

Table 8: Cross-patient trajectory variance decomposition. Each trajectory is fitted with a QR-decomposed polynomial basis (intercept, linear slope, quadratic curvature), and the variance of each coefficient across patients is expressed as a percentage of total cohort P/F variance. The intercept (baseline severity) accounts for ∼50% of total variance (the "severity trap") and is omitted from the table; columns show the remaining components. TC3 has a lower baseline component (37%) than other classes, meaning more of its total variance is attributable to trajectory dynamics, consistent with its non-monotonic biphasic shape.

Subgroup	**Slope**	**Curvature**	**Residual**
Overall	16.2%	6.1%	27.4%
TC1 (Early recovery)	22.4%	3.9%	26.0%
TC2 (Stable)	16.6%	5.3%	33.4%
TC3 (Biphasic)	17.7%	4.0%	41.3%
TC4 (Rapid decline)	14.9%	4.6%	20.8%

Slope plus curvature together contribute ∼22% of total cohort P/F variance: sufficient signal for shape-aware models to exploit, if they can disentangle it from the dominant baseline severity component (∼50%).

Appendix J. Cohort Characteristics

Table 9 provides cohort descriptive characteristics.

Table 9: Baseline characteristics and outcomes. Values are median [IQR] unless stated otherwise. *Age in ENC reported as median (IQR) using a 10-year bin format. **Norepinephrine equivalent: median dose among vasopressor recipients.

Variable	MIMIC-IV	EUKC	ENC
Cohort size			
n	3,938	2,888	3,592
Demographics			
Age, years*	64 [53–74]	61 [50–71]	60–69 (50–79)
Male sex, n (%)	2,408 (61.1%)	1,888 (65.4%)	2,251 (62.6%)
Outcomes			
ICU mortality, n (%)	1,270 (32.2%)	972 (33.7%)	941 (26.2%)
ICU length of stay, days	10 [6–16]	15 [8–27]	12 [7–22]
Organ support			
Invasive mech. ventilation, n (%)	3,183 (80.8%)	1,982 (68.6%)	3,077 (85.6%)
Vasopressor use, n (%)	1,750 (44.4%)	1,653 (57.2%)	2,728 (75.9%)
Norepinephrine equiv., μg/kg/min**	0.18 [0.08–0.58]	0.12 [0.07–0.24]	0.09 [0.05–0.13]
Respiratory parameters			
PaO_2/FiO_2, mmHg	188 [148–234]	195 [141–248]	225 [185–262]
$PaCO_2$, mmHg	41 [36–46]	40 [36–45]	41 [37–46]
PEEP, cmH_2O	8 [6–11]	8 [6–10]	11 [9–14]
Minute ventilation, L/min	9.8 [8.3–11.6]	9.2 [7.7–10.9]	9.1 [7.7–10.8]
Driving pressure, cmH_2O	15 [12–18]	15 [12–18]	17 [13–21]

Appendix K. External Validation for All Models

Table 10 and Figure 11 extend the main external-validation analysis (Section 6.3) to all seven primary models. Frozen MIMIC-IV–trained encoders were applied to EUKC and ENC; GMM-full (K=4) was re-fitted on each target cohort. The MIMIC-IV column reports the internal 3-seed values for direct comparison with the 3-seed external cohorts; these differ slightly from the 20-seed primary benchmark (Table 2; e.g. Model B 0.344 here vs. 0.330 over 20 seeds).

Native versus ported reference labels. The external TC_{ref} labels use the MIMIC-trained CRLCMM *ported* to each cohort, which could let a VAE inherit the reference's biases. We therefore re-evaluated the frozen encoders against *natively*-fit CRLCMMs at each site (de-novo four-class fits from the companion study). The native and ported references differ substantially (between-reference class ARI 0.59 at ENC, 0.44 at EUKC), yet re-scoring against the native labels leaves the picture essentially unchanged (e.g. ENC Model B 0.298 → 0.260, D3 0.302 → 0.285; EUKC flat to slightly higher). In the lower-mortality ENC cohort the native fit's parsimonious optimum is K=5; we retain K=4 throughout for cross-cohort comparability.

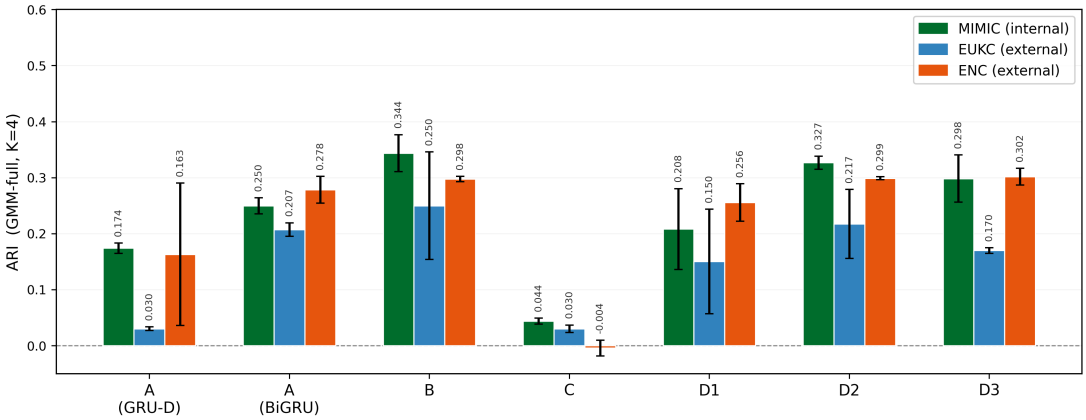


Figure 11: TC_{ref} recovery ARI (mean ± SD, 3 seeds, GMM-full, K=4) for all seven primary models across MIMIC-IV (internal), EUKC (external), and ENC (external). The MIMIC-IV internal ranking is broadly preserved externally. Model C fails on ENC (ARI = −0.004); Model A (GRU-D) collapses on EUKC (ARI = 0.030), consistent with its severity-dominated latent space being misaligned with EUKC's different severity distribution.

Table 10: TC_{ref} ARI (mean ± SD over 3 seeds, GMM-full, K=4) for all seven primary models on MIMIC-IV (internal test set), EUKC, and ENC. Bold: highest ARI per cohort column. †: negative ARI indicates clustering worse than random chance.

Model	**MIMIC-IV (internal)**	**EUKC (external)**	**ENC (external)**
A (GRU-D)	0.174±0.009	0.030±0.003	0.163±0.127
A (BiGRU)	0.250±0.014	0.207±0.012	0.278±0.024
B	**0.344±0.033**	**0.250±0.096**	0.298±0.005
C	0.044±0.005	0.030±0.007	−0.004±0.014 †
D1	0.208±0.072	0.150±0.093	0.256±0.033
D2	0.327±0.012	0.217±0.062	0.299±0.003
D3	0.298±0.042	0.170±0.005	**0.302±0.015**

The internal ranking (B > D2 > D3 > A BiGRU > D1 > A GRU-D > C) is broadly preserved at ENC but is compressed at EUKC, where all models show reduced ARI. Model B achieves the highest EUKC ARI; on ENC, B, D2, and D3 perform comparably. The collapse of Model C at ENC (ARI = −0.004) confirms that its paradoxical TC3 recall is an artefact of MIMIC-IV–specific cluster-size imbalance rather than a transferable shape representation. Model A (GRU-D) collapses at EUKC (ARI = 0.030, vs. 0.174 internally), consistent with its latent space being dominated by a severity axis that shifts under distribution change.

Appendix L. High-Certainty TC_{ref} Sensitivity Analysis

The CRLCMM trajectory class assignments used as our benchmark reference are probabilistic: each patient receives a posterior distribution over the four trajectory classes (TC1–TC4), and we use the hard maximum-posterior assignment as the reference label. Patients with low max-posterior probability may receive ambiguous labels, potentially inflating both correct and incorrect cluster matches and adding noise to ARI estimates.

To assess benchmark robustness, we repeated the primary GMM-full K=4 TC_{ref} recovery evaluation restricted to the 316 test-set patients (40%) whose maximum CRLCMM posterior probability exceeded 0.90. The median max-posterior across all patients is 0.847, so this threshold retains approximately the top 40% by label certainty. Table 11 reports full-cohort and high-certainty ARI for all seven models.

Table 11: High-certainty TC_{ref} sensitivity analysis. ARI restricted to test-set patients with CRLCMM max posterior probability > 0.90 (N= 316, 40% of test set; full-cohort N= 788). Full-cohort ARI shown at the original three-seed protocol for comparison (cf. the 20-seed primary benchmark, Table 2). HC ARI reported as mean ± SD over 3 seeds.

Model	Full-cohort ARI	HC ARI (N=316)
B	0.344 ± 0.033	0.626 ± 0.100
D2	0.327 ± 0.012	0.728 ± 0.018
D3	0.298 ± 0.042	0.646 ± 0.057
D1	0.208 ± 0.072	0.407 ± 0.056
A (BiGRU)	0.250 ± 0.014	0.457 ± 0.016
A (GRU-D)	0.174 ± 0.009	0.361 ± 0.170
C	0.044 ± 0.005	0.277 ± 0.009

All seven models achieve substantially higher ARI on the high-certainty subset (e.g. B: 0.344 → 0.626; D2: 0.327 → 0.728; C: 0.044 → 0.277), consistent with cleaner reference labels amplifying the recoverable signal. The relative model ordering is largely preserved: C remains weakest, D2 and D3 lead among constrained models. D2 overtakes B on the high-certainty subset (0.728 vs. 0.626). This preserved ranking confirms that the benchmark differences reflect genuine variation in trajectory-structure recovery rather than differential sensitivity to ambiguous reference assignments.

On the high-certainty subset, D2 (SSM+CR) achieves the highest ARI (0.728±0.018), surpassing Model B (0.626±0.100). This reversal suggests that D2's state-space decoder captures trajectory shape effectively among clearly typed patients, but its full-cohort performance is attenuated by borderline cases where severity dominates cluster assignment.
The relative model ordering is preserved across certainty levels, though absolute recovery remains limited and improves with label certainty.

Appendix M. Multivariate Feature Ablations

To test whether the severity trap is an artefact of univariate input, we trained all seven models on three additional feature sets: PF+PEEP (F=2), PF+$PaCO_2$ (F=2), and

PF+PEEP+$PaCO_2$ (F=3), using the same hyperparameters and seeds as the primary analysis. Daily PEEP and $PaCO_2$ were extracted from the same 3,938 MIMIC-IV stays (observed-day missingness: 42.5% for PEEP, 47.3% for $PaCO_2$). TC_{ref} was derived from P/F alone, so ARI decline with multivariate inputs partly reflects distributional shift.

Table 12: Severity leakage with multivariate inputs. Split-z models: R^2_{shape} (mean ± SD, 3 seeds). Single-z models: R^2_{full}. PF-only values from Table 6.

Model	PF only	PF+PEEP	PF+CO_2	PF+PEEP+CO_2
Single-z models (R^2_{full})				
A (GRU-D)	0.27	0.34	0.30	0.35
A (BiGRU)	0.36	0.38	0.27	0.33
B	0.44	0.43	0.44	0.47
Split-z models (R^2_{shape})				
C	0.29	0.35	0.28	0.34
D1	0.01	0.02	0.05	0.02
D2	0.14	**0.71**	**0.68**	**0.41**
D3	0.14	0.20	0.13	0.12

Table 13: TC_{ref} ARI with multivariate inputs (mean ± SD, 3 seeds, GMM-full K=4).

Model	PF only	PF+PEEP	PF+CO_2	PF+PEEP+CO_2
A (GRU-D)	0.174	0.021	−0.005	−0.015
A (BiGRU)	0.250	0.220	0.107	0.128
B	0.344	0.262	0.262	0.232
C	0.044	0.001	0.005	0.006
D1	0.208	0.076	0.117	0.090
D2	0.327	0.140	0.144	0.154
D3	0.298	0.205	0.242	0.238

Additional severity-correlated variables deepen rather than dissolve the severity trap. D2 is the most vulnerable: adding PEEP increases R^2_{shape} fivefold (from 0.14 to 0.71). D3's anchoring mechanism maintains low leakage throughout (0.12–0.20), confirming that the constraint generalises to multivariate settings. D1 achieves the lowest leakage (0.01–0.05) but consistently lower ARI than D3, reinforcing that disentanglement is not sufficient. ARI declines for all models, as expected given that TC_{ref} was derived from P/F alone; A (GRU-D) drops to negative ARI with PF+PEEP+CO_2, indicating that additional severity-correlated inputs actively pull the representation away from trajectory shape.

Clinical-severity covariates as probe targets. Probing the frozen latents against clinical-severity variables (per-stay means) shows the trap is primarily a *P/F*-severity trap with modest spillover: the latent encodes day-0 P/F strongly (R^2=0.23–0.45) and a composite clinical-severity index (SOFA components, lactate, vasopressor dose) only moderately ($R^2\approx$0.17–0.24), with individual non-P/F markers weaker still (lactate $R^2\leq$0.13; norepinephrine-equivalent ≈0.09). Charlson comorbidity was excluded as a static admission covariate.

Clinical-severity covariates as inputs. Adding daily clinical-severity channels as model *inputs* (lactate, FiO_2, norepinephrine rate, respiratory rate; F=3, 5, 7) deepens the trap in unconstrained models: shape-subspace leakage rises to R^2≈0.31–0.39 (Model C 0.28 → 0.37), while only D1's shape subspace keeps leakage near zero ($\leq$ 0.03 through F=7); D3 stays $\leq$ 0.13. (The linear state-space decoder of D2 is numerically unstable on the zero-inflated vasopressor input and is reported on the vasopressor-free sets only.) Richer severity-correlated inputs therefore worsen, not relieve, the trap unless the shape subspace is explicitly anchored.

Appendix N. MIMIC-IV v3.1 Replication

To confirm the findings are not specific to the legacy MIMIC-IV v1.0 extract, we re-derived the AHRF cohort from the current PhysioNet release, MIMIC-IV v3.1, using identical inclusion criteria and the 14-day window. This yields N=5,004 stays (∼97% overlap with the v1.0 cohort plus ∼1,170 new stays). TC_{ref} labels for v3.1 were assigned by applying the existing MIMIC CRLCMM without refitting (a zero-iteration update), giving a class distribution (889/2,187/912/1,016) close to v1.0. We retrained all seven models on v3.1 (single seed) and repeated the severity probe and TC_{ref} recovery. The severity trap reproduces: the severity probe $R^2(z \to$ day-0 P/F) is near-identical to v1.0 for every model (Table 14; e.g. Model B 0.510 vs 0.438, D1 0.232 vs 0.230), and the same model dissociations hold (D1 retains the lowest leakage and contamination). Absolute recovery ARIs track the v1.0 values as *magnitudes* (Pearson r=0.73 across the seven models); we report this as a value correlation, not a rank correlation; at this single seed the ordering reshuffles (Spearman ρ≈0.32), so we do not claim rank-preservation. The trap is therefore not an artefact of the v1.0 data version.

Table 14: MIMIC-IV v3.1 replication (N=5,004, single seed). Severity probe $R^2(z \to$ day-0 P/F) (v3.1 vs v1.0) and TC_{ref} ARI on v3.1. The severity probe is near-identical across versions.

Model	R^2 (v3.1)	R^2 (v1.0)	ARI (v3.1)
A (GRU-D)	0.293	0.273	0.333
A (BiGRU)	0.368	0.364	0.191
B	0.510	0.438	0.302
D2	0.487	0.440	0.297
D3	0.498	0.441	0.303
D1	0.232	0.230	0.206
C	0.348	0.313	0.075

Appendix O. Trial-Enrichment Simulation

To quantify the clinical cost of the severity trap we simulate a deterioration-enriched trial: patients predicted to be in a deteriorating class (TC3∪TC4) are enrolled, a treatment with a fixed relative risk reduction acts only on truly deteriorating patients (truth = CRLCMM class), and we compute the per-arm sample size for 80% power (two-proportion test) under each enrichment source (Table 15). This is *prognostic* (sample-size) enrichment only; it is *not* a treatment-effect-heterogeneity claim, which retrospective data cannot support under

confounding by indication. Enriching with the oracle CRLCMM day-3 probabilities needs 159 patients/arm; every feasible alternative: deep Model B clusters (using *full-trajectory* information, an optimistic upper bound), engineered day-0–2 features, or a day-2 P/F threshold, needs 3–6× more (483, 560, 860), and the deep representation reaches only ∼52% deteriorator purity. The severity-aligned geometry does not enable efficient enrichment, and is no better than engineered features at doing so.

Table 15: Prognostic trial-enrichment cost (MIMIC test; relative risk reduction 0.25 on true deteriorators; single seed). Per-arm N for 80% power and achieved deteriorator purity, by enrichment source.

Enrichment source	**Per-arm N**	**Purity**
Oracle CRLCMM (day-3)	159	1.00
Deep Model B clusters (full-traj.)	483	0.52
Engineered features (day 0–2)	560	0.59
Day-2 P/F threshold	860	—
No enrichment	919	—

Appendix P. Non-VAE Architecture Details

To test whether the severity trap is specific to variational autoencoding or is a more general property of the training objective (Section 6.4), we evaluated three additional families of sequential encoder using the same MIMIC-IV train/test split, masking scheme, robust scaling, and post-hoc GMM-full (K=4) evaluation as the VAEs. All are trained on the univariate P/F sequence; the frozen embedding is clustered and scored against $\mathrm{TC}_{\mathrm{ref}}$. Values are means over seeds 101–103. **Contrastive encoder (SimCLR-style).** A BiGRU encoder (hidden 64, latent z=32) trained with an NT-Xent instance-discrimination loss (τ=0.2, 150 epochs). Positives are generated by stochastic augmentation: jitter (σ=9 mmHg), magnitude warping, and random time-masking. To probe level-invariance we optionally add a *baseline-shift* augmentation that adds a constant $c \sim \mathcal{U}(-s, +s)$ to the whole series, with half-range $s \in \{0, 150, 250\}$ mmHg (s=0 is the standard, level-preserving contrastive baseline). Severity encoding falls monotonically with the shift range: $R^2(z \to \text{day-0 P/F}) = 0.973$ (s=0), 0.468 (s=150), 0.369 (s=250); TC2/TC3 ARI $= 0.188, 0.226, 0.213$ respectively. A stronger variant adding level-matched hard negatives and time-reversal negatives raises TC2/TC3 ARI to 0.263, still below the engineered baseline (0.43).

Masked-reconstruction transformer. A Transformer encoder (model dimension 64, 4 heads, 2 layers, learned positional encodings) trained to predict a held-out 30% of observed days (masked-value prediction); the latent is the pooled encoder output (dim 32), 150 epochs. ARI (all) = 0.302±0.018, TC2/TC3 ARI = 0.273±0.059, severity R^2 = 0.398±0.025.

Diffusion autoencoder. A BiGRU semantic encoder (z=16) conditioning a DDPM denoiser (100 steps, MLP) over the 14-day trajectory (200 epochs); we probe the semantic latent. Severity R^2 = 0.996–0.998 (the strongest severity dominance of any architecture tested); TC2/TC3 ARI = 0.08–0.38 across seeds.

Appendix Q. Supplementary Figures and Tables

For length, the following figures/tables are placed here and referenced from the main text: cluster trajectory profiles (Figure 12), external-validation transfer (Figure 13), per-model native-K recovery (Table 16), and external validation of Model D3 (Table 17).

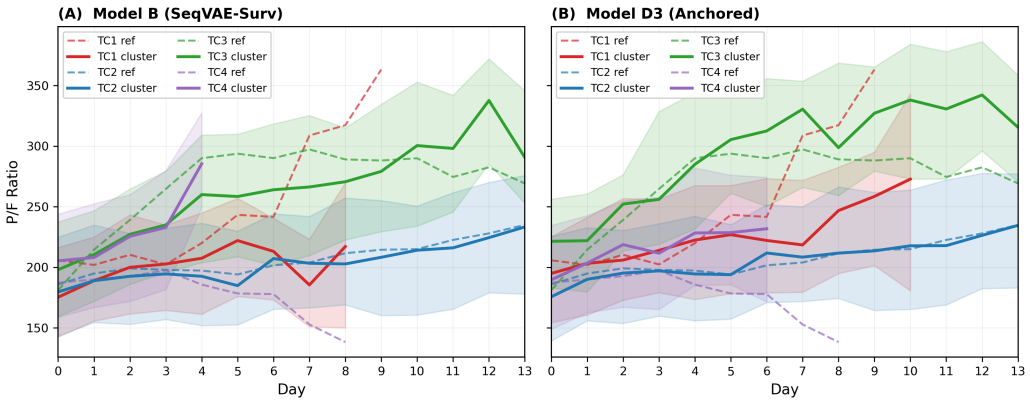


Figure 12: Cluster median P/F trajectories (solid, IQR shaded) overlaid on TC_{ref} medians (dashed) for **(A)** Model B and **(B)** Model D3; days with fewer than 10 observed patients are suppressed. Both recover the broad severity ordering, but D3's clusters more closely track the TC_{ref} trajectory shapes, including the biphasic TC3 rise-then-fall.

Table 16: Per-model native-K recovery (GMM-full at each model's BIC-optimal $K^\star$; test split, seeds 101–103). NMI and ARI are computed against the four-class TC_{ref} (both tolerate $K \neq 4$). η^2(level): fraction of baseline-severity variance explained by cluster membership. No model recovers shape at its native $K^\star$; the native clusters track severity level, and Model A (GRU-D)'s degenerate K=2 split is a severity dichotomy.

Model	$K^\star$	**NMI**	**ARI@$K^\star$**	η^2(level)	K=2 **sev. AUC**
D2	6	0.310	0.272	0.191	—
B	5	0.304	0.270	0.220	—
A (BiGRU)	3	0.234	0.241	0.484	—
D1	4	0.299	0.220	0.188	—
D3	7	0.296	0.213	0.406	—
A (GRU-D)	2	0.081	0.050	0.485	**0.958**
C	5	0.250	0.039	0.313	—

Table 17: External validation of Model D3 (frozen MIMIC encoder applied to external cohorts). Mean ± SD over 3 seeds. $R^2(z_s)$: severity leakage into the shape subspace; higher values at external sites indicate partial loss of disentanglement under distribution shift.

Cohort	ARI	NMI	Macro-F1	$R^2(z_s)\downarrow$	$R^2(z_\ell)$
MIMIC (internal)	0.298±0.042	0.312±0.021	0.488±0.035	0.140±0.051	0.258±0.011
EUKC (external)	0.170±0.005	0.266±0.001	0.407±0.007	0.348±0.113	0.449±0.005
ENC (external)	0.302±0.015	0.352±0.008	0.431±0.024	0.272±0.079	0.358±0.006

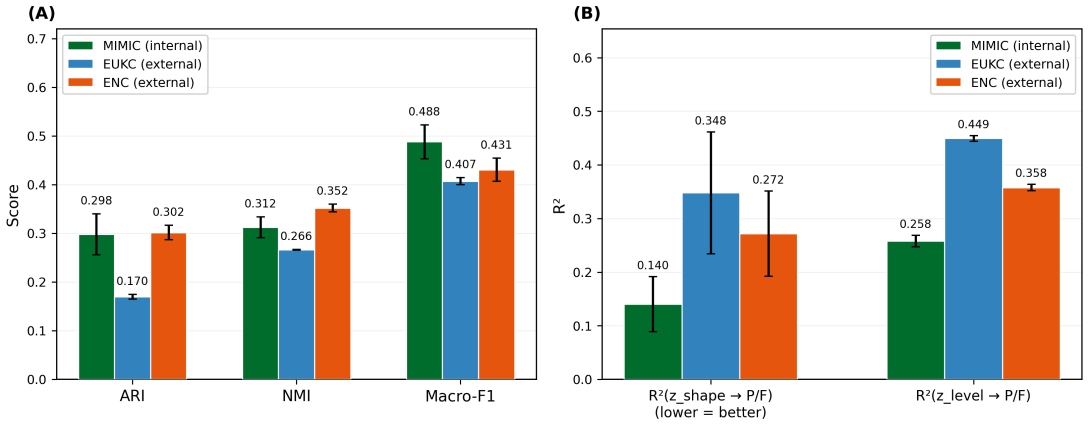


Figure 13: External validation of Model D3 across the derivation cohort (MIMIC) and two external cohorts (EUKC, ENC); frozen MIMIC-trained encoder, GMM-full K=4, mean ± SD over 3 seeds. **(A)** TC_{ref} recovery (ARI, NMI, macro-F1): external recovery matches internal for ENC and is modestly lower for EUKC. **(B)** Severity leakage, R^2 of each latent subspace predicting baseline P/F: leakage into the shape subspace (lower is better) stays modest internally (0.14) and rises externally (EUKC 0.35, ENC 0.27), indicating partial loss of disentanglement under distribution shift.

Data & Code Availability

MIMIC-IV data are available at https://physionet.org/content/mimiciv/. Code is available at https://github.com/ai4ai-lab/Severity_Trap.

Ethics and Data Governance

MIMIC-IV is a publicly available de-identified database approved by the Beth Israel Deaconess Medical Center (BIDMC) IRB (protocol 2001-P-001699/14). Access was obtained via PhysioNet credentialing in accordance with the PhysioNet Credentialed Health Data Use Agreement. The EUKC dataset was approved by the local research ethics committee. The ENC dataset was approved under the institutional data governance framework. No additional

IRB review was required for this secondary analysis of pre-existing anonymised cohorts. Informed consent was waived by the approving bodies given the retrospective, de-identified nature of the data.