DEEP VARIATIONAL FRAILTY MODELS FOR SURVIVAL PREDICTION UNDER DEPENDENT CENSORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Most survival models assume that censoring is uninformative about event time given observed covariates, an assumption that fails when unobserved factors drive both event and censoring times. Such dependence cannot be identified from right-censored data without structural assumptions. We propose the Deep Variational Frailty Model (DVFM), which jointly models event and censoring times through a shared latent frailty and nonlinear conditional decoders rather than a prespecified copula family. DVFM is trained with a β-regularized censored-data variational objective and amortized inference. DVFM is neither an assumption-free identification method nor a test for dependent censoring, but a predictive model for settings where shared frailty plausibly explains the dependence. On synthetic and semi-synthetic benchmarks, DVFM matches the strongest dependent-censoring baseline and a Cox-Gamma frailty model on IBS and MAE, and outperforms four frailty-free models that assume independent censoring by 21-35% within a dataset. It also recovers subject-level frailty better than scalar frailty baselines, mainly on censored subjects, whose censoring times other models treat as uninformative. A matched ablation without the latent (d_z=0) shows that the latent drives the gains under dependence in synthetic data, while on semi-synthetic data its benefit is dataset-dependent: it lowers IBS on seven of twelve datasets, by 10-25% on four, and raises it by about 8% on two. On real data, where we predict time to death in ALS patients, the latent improves held-out log-likelihood.

1 INTRODUCTION

Survival analysis models time-to-event outcomes, such as relapse or death in healthcare, failure in engineering, or default in finance. This is often complicated due to *censoring*: the study ends or a subject withdraws before the event of interest, so we know only a lower bound on this time. Standard estimators and most deep learning-based survival models assume *independent censoring*: given covariates $\boldsymbol{X}$, the event time E and the censoring time C are independent. When this fails, censoring is considered *dependent*, and a common cause is an unobserved factor that is relevant to both processes. Suppose patients with high-grade tumors relapse sooner, but they also leave follow-up sooner, because they become too ill to attend visits, move to palliative care, or transfer to specialist centres. Advanced disease then shortens both the time to relapse and the time to withdrawal, if a clinical study does not record "Tumor grade", then event times E and censoring times C are dependent given the covariates we do observe.

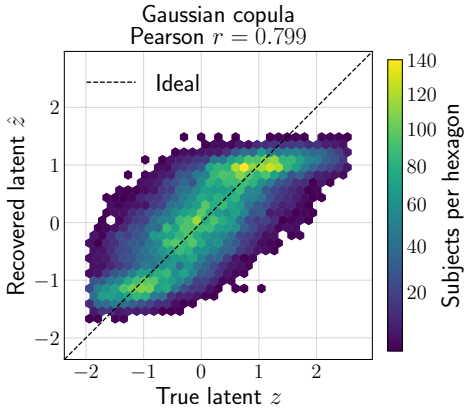


Figure 1: Recovered frailty $\widehat{z}$ (unobserved subject-level risk factor driving event and censoring times) against true z on held-out SUPPORT subjects. Gaussian copula, $\tau = 0.5$, 32% censoring; 10 repeats of 1,775 subjects.

Dependent censoring complicates the analysis for a more basic reason: the data alone cannot reveal it. For each subject, we see only whichever comes first, the event or the censoring, never both. Tsiatis (1975) proved formally that this means that any

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.

world in which the two are linked, can be matched by one in which they are independent, as it will produce the same distribution of observations, but with a different event-time distribution. Without further assumptions, no amount of data can tell the two apart– a problem known as *non-identifiability*. Identification can be restored only by structural assumptions, such as proportional-hazards regressors with sufficient covariate variation (Heckman & Honoré, 1989; Abbring & Van Den Berg, 2003). The standard answer so far has been a copula: a parametric dependence family is imposed between the event and censoring marginals, and recent deep learning methods either fix that family in advance or learn within one – see examples by Foomani et al. (2023); Zhang et al. (2024); Liu et al. (2025); Lillelund et al. (2026b).

We take the opposite route and instead model dependence through a shared latent variable. The unmeasured predictive variable of our opening example (i.e. "Tumor grade") is what the survival literature calls a *frailty*: an unobserved, subject-specific random effect acting multiplicatively on the hazard, $h(t \mid z, \boldsymbol{x}) = z\, h_0(t) \exp(\boldsymbol{\beta}^\top \boldsymbol{x})$ (Vaupel et al., 1979). Normally, a frailty is shared across a cluster of failure times that are conditionally independent given z but dependent once z is marginalized out (Clayton, 1978); the event and censoring times of a single individual form exactly such a pair, which is the mixed proportional hazards competing-risks model of econometrics (Abbring & Van Den Berg, 2003). Once the model is written as $p(e, c, z \mid \boldsymbol{x}) = p(e \mid \boldsymbol{x}, z)\, p(c \mid \boldsymbol{x}, z)\, p(z)$, we no longer choose the dependence at all – the shared z decides it. Figure 1 shows the frailty our model infers for held-out SUPPORT subjects against the latent that generated their data: the two track the identity line (Pearson $r = 0.80$), so the common cause is recovered per subject.

We formulate this as the **Deep Variational Frailty Model (DVFM)**, which pairs nonlinear conditional event and censoring decoders with amortized variational inference over the shared latent. DVFM replaces one structural assumption with another and our experiments ask whether that approach leads to better results. Our contributions are:

1. **Model.** We introduce DVFM, a deep generative model of $p(E, C \mid \boldsymbol{X})$ in which a shared subject-level latent frailty Z induces event-censoring dependence. This replaces the choice of a copula family with a mechanism learned from the data. It also replaces a single global dependence parameter with a posterior over each subject's frailty, so the model reports who it infers to have been informatively censored, and how confidently.
2. **Variational objective.** We derive the censored-data ELBO and distinguish it from the β-regularized training objective, which is not an evidence lower bound when $\beta \neq 1$.
3. **Theory and scope.** We give two results. First, under a restricted multiplicative-frailty model with known hazard indices and sufficient covariate variation, we establish population identification of the frailty-effect distribution. Second, we show that the population ELBO does not separate observationally equivalent frailty and independent-proxy candidates. These results characterize the model's inductive bias, but do not claim unrestricted identification of the full neural DVFM.
4. **Empirical study.** Among censored subjects, whose censoring times other models treat as uninformative, DVFM recovers the generating frailty better than scalar frailty baselines (Spearman 0.77-0.83 vs. 0.58-0.66 for Cox-Gamma on three SEER cohorts). On prediction it matches the strongest baselines: across 12 semi-synthetic datasets and four copula families ($\tau = 0.5$), it ranks second of nine models on IBS and MAE, and has slightly lower joint event-censoring error than HACSurv, though not significantly ($p = 0.12$). The latent improves prediction above Kendall's $\tau \approx 0.25$ on synthetic data and raises held-out likelihood in every seed on real ALS data. DVFM thus recovers *who* is frail but not *how much* dependence there is: the learned τ moves in the right direction but is not yet calibrated across datasets.

2 Background and Related Work

2.1 Preliminaries and problem definition

We consider a survival dataset $\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$ denotes covariates, $t_i \in \mathbb{R}_+$ the observed follow-up time, and $\delta_i \in \{0, 1\}$ the event indicator. Each instance has latent event and censoring times e_i and c_i, of which we observe only $t_i = \min(e_i, c_i)$ and $\delta_i = \mathbb{1}(e_i \leq c_i)$. At the population level, E and C are non-negative random variables with conditional survival functions $S_E(\, e \,|\, \boldsymbol{x}) = \Pr(\, E > e \,|\, \boldsymbol{x})$ and $S_C(\, c \,|\, \boldsymbol{x}) = \Pr(\, C > c \,|\, \boldsymbol{x})$, and $T = \min(E, C)$ and

$\Delta = \mathbb{1}[E \leq C]$ denote the observed follow-up time and event indicator, with realizations t_i and δ_i. Our target is $S_E(\, e \,|\, \boldsymbol{x})$, including for subjects whose follow-up ends before the event.

Most survival methods assume independent censoring, $E \perp\!\!\!\perp C \mid \boldsymbol{X}$ (Kleinbaum & Klein, 2012); censoring is dependent when $E \not\perp\!\!\!\perp C \mid \boldsymbol{X}$ (Appendix B). Without further assumptions, the event-time distribution is not identifiable from right-censored data (Tsiatis, 1975). Regressors entering the event and censoring risks differently can restore identification within structured model classes (Heckman & Honoré, 1989; Abbring & Van Den Berg, 2003), as assumed by our identification result (Section 4, Appendix E). DVFM assumes

$$E \perp\!\!\!\perp C \mid (\boldsymbol{X}, Z), \tag{1}$$

which allows dependence given $\boldsymbol{X}$ through the shared latent variable Z. This restricts the model class; it does not eliminate general non-identifiability.

2.2 Related work

Survival models under independence. CoxPH (Cox, 1972), RSF (Ishwaran et al., 2008), DeepSurv (Katzman et al., 2018), and discrete-time models such as MTLR (Yu et al., 2011) span classical, ensemble, and neural approaches, but agree on the treatment of censoring: a subject censored at time t either drops out of the risk set or contributes the survival probability $S_E(\, t \,|\, \boldsymbol{x})$, both of which assume conditionally independent censoring given $\boldsymbol{x}$.

Copula approaches to dependent censoring. Foomani et al. (2023) pair deep marginals with a copula family fixed in advance, so the dependence geometry enters as a modeling input rather than an estimand. DCSurvival (Zhang et al., 2024) removes that input by learning an Archimedean generator end-to-end alongside the marginals, with identifiability under parametric marginals; HACSurv (Liu et al., 2025) generalizes to hierarchical Archimedean copulas for asymmetric dependence among competing events and censoring. The progression relaxes *which* copula but not the stance: dependence is a coupling between two marginals, constrained to a family, with a generator global across subjects. DVFM represents it instead through an inferred subject-level latent, allowing dependence to vary with covariates and making the latent itself an object of inference.

Frailty models. Shared frailty induces dependence among conditionally independent time-to-event processes through a random effect (Vaupel et al., 1979; Hougaard, 1986; Oakes, 1989; Duchateau & Janssen, 2007). Cox-Gamma (CG) admits closed-form subject-level inference but assumes Gamma frailty, proportional hazards, and multiplicative effects. The Neural Frailty Machine (NFM) (Wu et al., 2023) relaxes the functional form, but applies frailty to the event process alone, retains the independent-censoring likelihood, and estimates the frailty distribution rather than an individual posterior. The closest precedent is Huang & Wolfe (2002), where a cluster-level frailty enters both the failure and censoring hazards, identified by within-cluster replication. DVFM moves this construction to the subject level, where no replication exists, with a neural decoder and amortized per-subject inference; Section 4 gives identification conditions without clustering.

Deep latent-variable survival models. VSI (Xiu et al., 2020) models $p(E \mid \boldsymbol{X})$ using a time-informed posterior, while SAVAE (Apellániz et al., 2024) and VaDeSC (Manduchi et al., 2022) generate covariates and event times from a shared latent variable, VaDeSC with a mixture prior that clusters subjects by survival mechanism. None models the censoring time C, and thus none represents event-censoring dependence. CDVI (Liu & Wang, 2025) improves latent inference by conditioning the variational posterior on the censoring indicator, but assumes $E \perp\!\!\!\perp C \mid \boldsymbol{X}$ and models only the event-time distribution. DVFM instead jointly generates E and C from a shared latent frailty and infers that frailty for held-out subjects.

3 Method: Deep Variational Frailty Model

DVFM assumes that dependence arises through a shared latent frailty variable $Z \in \mathbb{R}^{d_z}$ such that (1) holds. The joint generative model is

$$p_\theta(e, c, z \mid \boldsymbol{x}) = p_\theta(e \mid \boldsymbol{x}, z)\;\; p_\theta(c \mid \boldsymbol{x}, z)\;\; p(z), \tag{2}$$

with prior $p(z) = \mathcal{N}(0, I)$. The latent variable z represents unobserved heterogeneity such as latent health status, frailty, or failure propensity.

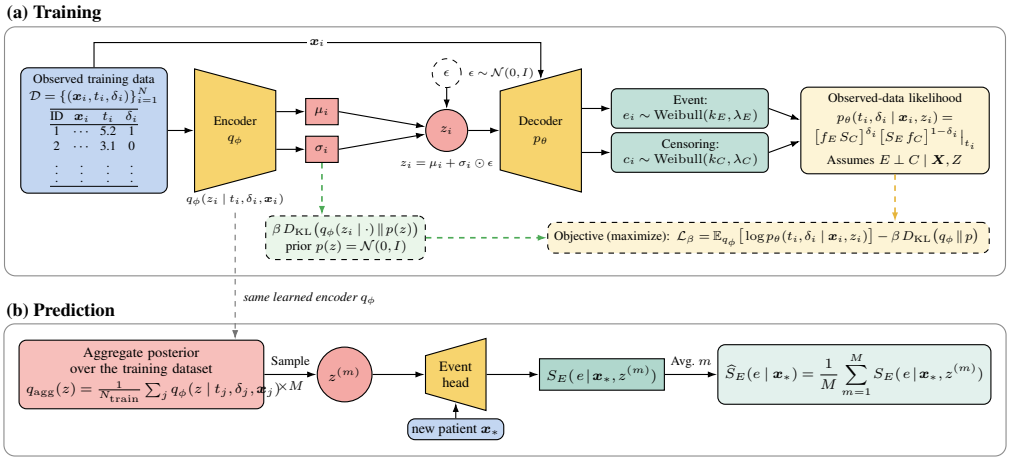


Figure 2: Training and prediction procedure of DVFM. (a) The encoder q_ϕ maps $(t_i, \delta_i, \boldsymbol{x}_i)$ to the parameters (μ_i, σ_i) of a Gaussian variational posterior, and the latent frailty is drawn by reparameterization, $z_i = \mu_i + \sigma_i \odot \epsilon$. The decoder p_θ receives $(\boldsymbol{x}_i, z_i)$ and outputs Weibull parameters $(k_E, \lambda_E, k_C, \lambda_C)$ for the latent event time e_i and censoring time c_i. Assuming E and C are conditionally independent given $(\boldsymbol{X}, Z)$, these distributions induce the likelihood of the observed pair (t_i, δ_i) used in the β-regularized variational objective. (b) For a new patient with covariates $\boldsymbol{x}_*$, the same encoder is applied to the training split to form the aggregate posterior q_{agg}; frailties $z^{(m)}$ are sampled from it, passed with $\boldsymbol{x}_*$ through the event head, and the resulting conditional survival curves $S_E(\, e \,|\, \boldsymbol{x}_*, z^{(m)})$ are averaged to obtain the estimate $\widehat{S}_E(e \mid \boldsymbol{x}_*)$ of the target $S_E(\, e \,|\, \boldsymbol{x}_*)$.

Induced dependence. We start by writing S_E, S_C for the conditional survival functions, which are induced by (2), and marginalizing over the latent gives

$$S(\, e, c \,|\, \boldsymbol{x}) \;=\; \int S_E(\, e \,|\, \boldsymbol{x}, z)\; S_C(\, c \,|\, \boldsymbol{x}, z)\; p(z)\; dz, \tag{3}$$

which need not factorize into $S_E(\, e \,|\, \boldsymbol{x})\, S_C(\, c \,|\, \boldsymbol{x})$. DVFM therefore induces dependence without fixing a copula family: the departure from independence is exactly a covariance over the latent.

Proposition 1 (Dependence induced by latent marginalization)**.** *Fix a covariate vector $\boldsymbol{x}$. Suppose* (1) *holds and that*

$$Z \mid \boldsymbol{X} = \boldsymbol{x} \sim p(z).$$

Then, for all $e, c \geq 0$, we have

$$S(e, c \mid \boldsymbol{x}) - S_E(e \mid \boldsymbol{x}) S_C(c \mid \boldsymbol{x}) = \text{Cov}_{Z \sim p(z)}\left(S_E(e \mid \boldsymbol{x}, Z), S_C(c \mid \boldsymbol{x}, Z)\right). \tag{4}$$

Consequently, E and C are dependent conditional on $\boldsymbol{X} = \boldsymbol{x}$ if and only if this covariance is nonzero for at least one pair (e, c). (Proof in Appendix D.)

Neural parameterization of event and censoring. A decoder network $[\psi_E(\boldsymbol{x}, z), \psi_C(\boldsymbol{x}, z)] = \text{Decoder}_\theta(\boldsymbol{x}, z)$ outputs the distributional parameters of the event and censoring models. In our implementation both times are conditionally Weibull, with shapes and scales that depend on $(\boldsymbol{x}, z)$:

$$E \mid \boldsymbol{x}, z \sim \text{Weibull}(k_E, \lambda_E), \qquad C \mid \boldsymbol{x}, z \sim \text{Weibull}(k_C, \lambda_C). \tag{5}$$

Positivity of the Weibull parameters is enforced by construction: the shape parameters use a softplus activation, and the scale parameters an exponential link, each offset by a 10^{-6} floor.

Amortized variational inference. The posterior $p(z \mid t, \delta, \boldsymbol{x})$ is intractable, so we introduce an encoder network defining a Gaussian variational posterior, given by:

$$q_\phi(z \mid t, \delta, \boldsymbol{x}) = \mathcal{N}\big(z; \mu_\phi(t, \delta, \boldsymbol{x}), \text{diag}(\sigma^2_\phi(t, \delta, \boldsymbol{x}))\big)\,. \tag{6}$$

We train the model using the reparameterization trick, writing $z = \mu_\phi + \sigma_\phi \odot \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$.

Censored variational objective. Let $Y = (T, \Delta)$, with realization $y = (t, \delta)$, so that the encoder of (6) may equivalently be written $q_\phi(z \mid y, \boldsymbol{x})$. Under (1), the observed-data likelihood contribution of one instance $(t, \delta, \boldsymbol{x})$ conditional on z is

$$p_\theta(t, \delta \mid \boldsymbol{x}, z) = \left[f_E(t \mid \boldsymbol{x}, z) S_C(t \mid \boldsymbol{x}, z)\right]^\delta \left[S_E(t \mid \boldsymbol{x}, z) f_C(t \mid \boldsymbol{x}, z)\right]^{1-\delta}. \tag{7}$$

For an amortized posterior $q_\phi(z \mid y, \boldsymbol{x})$, the ordinary censored-data ELBO is

$$\mathcal{L}_1(\theta, \phi; y, \boldsymbol{x}) = \mathbb{E}_{q_\phi(z \mid y, \boldsymbol{x})}[\log p_\theta(y \mid \boldsymbol{x}, z)] - D_{\mathrm{KL}}(q_\phi(z \mid y, \boldsymbol{x}) \| p(z)), \tag{8}$$

which satisfies the exact identity

$$\mathcal{L}_1(\theta, \phi; y, \boldsymbol{x}) = \log p_\theta(y \mid \boldsymbol{x}) - D_{\mathrm{KL}}(q_\phi(z \mid y, \boldsymbol{x}) \| p_\theta(z \mid y, \boldsymbol{x})), \tag{9}$$

and is therefore a lower bound on the observed-data log likelihood.

In practice, we optimize the β-regularized objective

$$\mathcal{L}_\beta(\theta, \phi; y, \boldsymbol{x}) = \mathbb{E}_{q_\phi(z \mid y, \boldsymbol{x})}[\log p_\theta(y \mid \boldsymbol{x}, z)] - \beta D_{\mathrm{KL}}(q_\phi(z \mid y, \boldsymbol{x}) \| p(z)). \tag{10}$$

Equivalently, $\mathcal{L}_\beta = \mathcal{L}_1 + (1 - \beta) D_{\mathrm{KL}}(q_\phi(z \mid y, \boldsymbol{x}) \| p(z))$. Thus, when $\beta \neq 1$, $\mathcal{L}_\beta$ is a regularized censored variational objective rather than an ordinary evidence lower bound. In particular, $\beta < 1$ can favor an informative encoder relative to the ordinary ELBO. Free-bits modifications also alter the objective and are used only as optimization regularizers. In all reported runs β is annealed to $\beta_{\max} = 1$ and free bits are disabled, so after warm-up the trained objective is $\mathcal{L}_1$ minus an ℓ_1 penalty, with weight α, on the decoder weights that carry z into the margin heads (Appendix G.7); the penalty is a regularizer and not part of the evidence lower bound.

Prediction via latent marginalization. For any new subject we observe only their covariates, so the predictive event-time survival function marginalizes the latent. DVFM predicts the aggregate posterior under the learned latent distribution:

$$q_{\mathrm{agg}}(z) = \frac{1}{N_{\mathrm{train}}} \sum_{i=1}^{N_{\mathrm{train}}} q_\phi(z \mid t_i, \delta_i, \boldsymbol{x}_i), \tag{11}$$

giving $S_E(e \mid \boldsymbol{x}) = \int S_E(e \mid \boldsymbol{x}, z)\, q_{\mathrm{agg}}(z)\, dz$, which we approximate by

$$\widehat{S}_E(e \mid \boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^{M} S_E(e \mid \boldsymbol{x}, z^{(m)}), \qquad z^{(m)} \sim q_{\mathrm{agg}}. \tag{12}$$

Because q_{agg} pools the training subjects' posteriors, prediction requires encoding the training split but uses no test-set information and no follow-up for the subject being predicted; prior sampling would be valid only if $q_{\mathrm{agg}} \approx p$. Architecture, objective weighting, optimization and numerical safeguards are given in Appendix G.7.

4 What DVFM Can and Cannot Identify

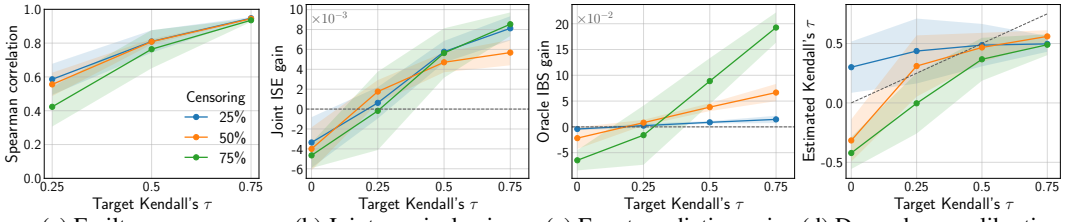
What can be identified under a restricted frailty structure? Although the joint distribution of (E, C) is not identifiable from right-censored data alone (Tsiatis, 1975), the frailty-effect distribution is identified within a restricted multiplicative-frailty subclass. Under three conditions – multiplicative frailty, $S_E(e \mid \boldsymbol{x}, z) = \exp\{-\Lambda_E(e, \boldsymbol{x})\, \eta(z)\}$ and $S_C(c \mid \boldsymbol{x}, z) = \exp\{-\Lambda_C(c, \boldsymbol{x})\, \zeta(z)\}$ for frailty effects $\eta(z), \zeta(z) > 0$; a frailty-effect law H common to all $\boldsymbol{x}$; and known cumulative hazards whose image $\{(\Lambda_E(t, \boldsymbol{x}), \Lambda_C(t, \boldsymbol{x}))\}$ contains a nonempty open set – the population law of $(T, \boldsymbol{X})$ identifies H, and hence the joint and marginal survival functions (Theorem 1, Appendix E). The proof expresses the observed survival function as the bivariate Laplace transform L_H evaluated on that image and extends by analyticity, following Abbring & Van Den Berg (2003), who use this argument to weaken the full-quadrant condition of Heckman & Honoré (1989) to an open set. Unlike them, we assume the hazards are known; in exchange, they need not be proportional in $(t, \boldsymbol{x})$. The identified object is the law of the frailty effects, not individual realized frailties, and the implemented decoder – whose Weibull shape depends on z – need not satisfy the multiplicative separation the result assumes: a latent that changes the shape of the hazard in time does not act as a multiplicative factor on it. Fixing the shape, a configuration our implementation supports, is necessary but not sufficient: with a shared shape k, $S_E(e \mid \boldsymbol{x}, z) = \exp\{-e^k \lambda_E(\boldsymbol{x}, z)^{-k}\}$ has the theorem's form only if the log-scale is additive, $\log \lambda_E(\boldsymbol{x}, z) = f(\boldsymbol{x}) + g(z)$, which a decoder over the concatenated $(\boldsymbol{x}, z)$ does not guarantee, and the theorem further takes the cumulative hazards as known.

What the variational objective can and cannot prefer? The role of the variational objective is best understood as an objective-comparison statement rather than a dependence-recovery theorem. Let $r \in \{f, p\}$ index a frailty-based and an independence-proxy candidate, and write $\ell_r = \mathbb{E}_{p_0} \log p_{\theta_r}(Y \mid \boldsymbol{X})$ for the expected observed-data log-likelihood, G_r for the expected posterior gap $D_{\mathrm{KL}}(q_{\phi_r} \| p_{\theta_r}(Z \mid Y, \boldsymbol{X}))$, and K_r for the expected $D_{\mathrm{KL}}(q_{\phi_r} \| p(Z))$. Substituting $\mathbb{E}\mathcal{L}_1 = \ell - G$ and $\mathcal{L}_\beta = \mathcal{L}_1 + (1-\beta)K$ gives

$$\mathbb{E}\mathcal{L}_\beta(\theta_f, \phi_f) - \mathbb{E}\mathcal{L}_\beta(\theta_p, \phi_p) = (\ell_f - \ell_p) - (G_f - G_p) + (1-\beta)(K_f - K_p). \tag{13}$$

At $\beta = 1$ the frailty candidate is preferred only when its likelihood advantage exceeds its excess variational gap; if the two candidates induce the same observed-data distribution and both exact posteriors lie in the variational family, the population ELBO ties them. The objective therefore does not select the true dependent decomposition among observationally equivalent models. Under independent censoring, if the true density belongs to a latent-inactive submodel, an optimum exists in which the posterior equals the prior and the latent carries no dependence information – an existence result only, since nonconvex optimizers may still use the latent. Under dependent censoring, DVFM can prefer the frailty explanation when it gives a sufficiently better fit after its variational gap and regularization cost. This is an inductive-bias argument, not unrestricted identification of (E, C); formal statements and proofs are in Appendix F.

5 Synthetic Experiments



(a) Frailty recovery (b) Joint-survival gain (c) Event-prediction gain (d) Dependence calibration

Figure 3: Mean ($\pm$ SD) over 10 repeats of $N = 10{,}000$ subjects with $d_x = 10$ covariates on synthetic data; color denotes the censoring rate. (a) Spearman correlation between recovered and true latent. (b) Joint-survival ISE gain over an otherwise identical model with the latent removed ($d_z = 0$), paired within each repeat; positive values favor DVFM. Gains are absolute differences in each metric's own units (d_z=0 minus d_z=1); relative gains quoted in the text divide this difference by the d_z=0 value. (c) Oracle IBS gain, paired as in (b); both gains cross zero at $\tau \approx 0.25$. (d) Estimated versus target Kendall's τ; the diagonal (black dashed) line is perfect calibration.

Setup. We first test whether DVFM recovers a known shared latent in fully synthetic data. We generate event and censoring times using Weibull predictors that depend on the same standard-normal subject frailty and $d_x = 10$ covariates (Appendix G.1). The construction is multiplicative in the frailty, so the data lie inside the restricted class of Theorem 1, while the fitted decoder does not, since its Weibull shape depends on z. We evaluate $\tau \in \{0, 0.25, 0.5, 0.75\}$, set by calibrating the shared-frailty loading, with censoring rates $\{0.25, 0.50, 0.75\}$ over 10 repeats of $N = 10{,}000$ subjects, split 70/10/20. Each dataset is fitted twice, with $d_z = 1$ and a matched $d_z = 0$ control (see Appendix G.7). All metrics are oracle metrics, computed on the test set against the true event times. See details in Appendix G.5.

Results. Three trends stand out. First, recovery improves with dependence. Figure 3 (a), shows that the more strongly a shared frailty drives both times, the more each subject's (t, δ) reveals about its z, and at $\tau = 0.75$ the latent is recovered almost exactly (Spearman ≈ 0.94) at every censoring rate. Under independence, adding the latent increases error, especially with heavy censoring. In Figures 3 (b) and (c) show that gains in joint ISE[1] and oracle IBS over $d_z = 0$. Under independence, it merely adds variance and a spurious link between the times, most costly under heavy censoring. Under dependence, a model without it treats censored subjects as representative of those still at risk, which

[1]The squared difference between the predicted and true joint survival surfaces $S(e, c \mid \boldsymbol{x})$, integrated over the (e, c) grid, normalized by its area, and averaged across subjects. See details in Appendix G.5.

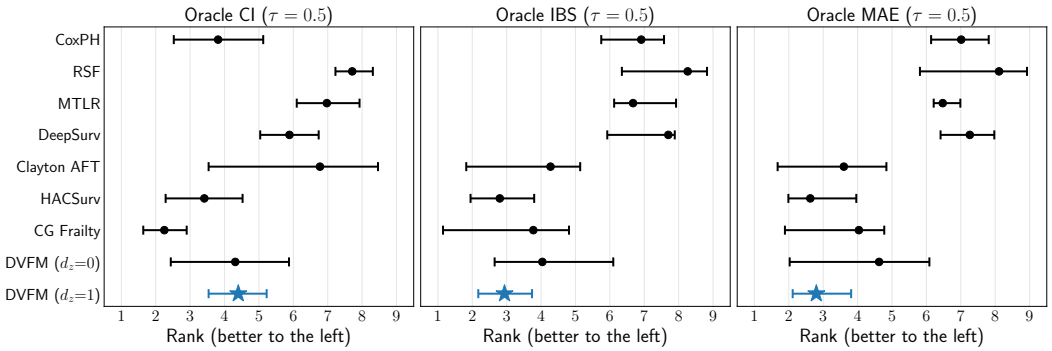


Figure 4: Prediction performance at Kendall's $\tau = 0.5$, as model ranks on oracle CI, IBS and MAE (rank 1 is best). We rank the nine models – the eight benchmarks and DVFM's matched d_z=0 ablation – within each dataset × copula × seed cell, then average over the 10 seeds and four copulas to give one mean rank per entry per dataset. Markers show the median of those 12 dataset values, with DVFM shown as a star; bars are 95% bootstrap percentile intervals over datasets (see Appendix G.6).

the latent variable corrects this; the gain grows with both dependence and censoring, and at strong dependence with heavy censoring the latent removes almost all joint-survival error. Third, the model detects dependence but does not measure it, as Figure 3 (d) shows: the learned τ rises with the target but is shrunk at strong dependence and is biased under independence, as Section 4 anticipates. The shared latent therefore pays off only when there is dependence to model.

6 Semi-synthetic experiments

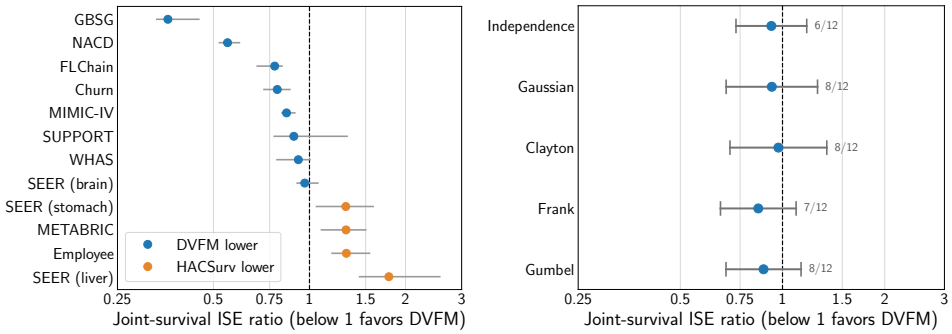
Setup. We now ask whether the shared latent still pays off on realistic covariate distributions, when the data come from a model outside DVFM's parametric family, and against copula baselines that fix the dependence family in advance (Clayton AFT) or learn a single Archimedean generator for the whole dataset (HACSurv). Real data cannot answer this: the true event time of a censored subject is never observed, so the oracle metrics are unavailable. Following Foomani et al. (2023, Algorithm 4), we keep a source dataset's covariates, fit event and censoring margins to it, and resample both times through a copula of known Kendall's τ, giving realistic covariates and marginals with a known event time for every subject. Appendix G.2 gives our construction, where it departs from theirs, and the three ways in which the benchmark lies outside DVFM's model class. We use 12 source datasets, four copula families (Gaussian, Clayton, Frank, Gumbel) at $\tau \in \{0, 0.5\}$, and 10 seeds, with every model sharing the dataset and split within a seed. We tune hyperparameters per dataset on a development split never reused in evaluation (see Appendices G.7 and G.8). We report oracle CI, IBS and MAE, see Appendix G.5. Appendix G.3 gives datasets and preprocessing.

Results – Predictive Performance. Figure 4 ranks the nine models at Kendall's $\tau = 0.5$. DVFM is second on oracle IBS and MAE, and fifth on oracle CI. Bootstrap intervals overlap on every metric, so the ranks summarize rather than separate, and the spread behind them is what matters. On IBS and MAE the four frailty-free independence baselines (CoxPH, DeepSurv, RSF, MTLR) are 21%-35% worse than DVFM within a dataset, while Cox-Gamma frailty stays within 1.5% (see Table 7). All oracle CIs fall within a two-percent band: the C-index reads only the ordering of predicted medians, which survives a badly scaled survival curve, so it cannot see the error dependence creates. DVFM's CI gap to the models above it is at most 0.006, below the 0.008 median seed-to-seed standard deviation per cell, so we do not report CI alone.

Table 1: Model rank at $\tau = 0 \rightarrow \tau = 0.5$, median over the 12 source datasets, DVFM at d_z=1; rank 1 is best, so a fall is an improvement.

Model	Oracle IBS	Oracle CI	Oracle MAE
Relaxes independent censoring			
DVFM	4.15 → 2.49	4.60 → 3.86	4.45 → 2.45
Clayton AFT	6.00 → 3.64	6.70 → 5.81	5.20 → 3.10
HACSurv	4.00 → 2.20	3.45 → 2.89	4.10 → 2.20
Assumes independent censoring			
RSF	6.90 → 7.30	6.10 → 6.82	6.70 → 7.19
MTLR	4.85 → 5.79	5.65 → 5.97	5.20 → 5.86
DeepSurv	5.05 → 6.70	4.80 → 5.19	4.85 → 6.44
CG Frailty	1.55 → 3.66	2.05 → 1.79	1.70 → 3.64
CoxPH	3.25 → 5.91	2.20 → 3.38	3.70 → 6.20

Table 1 reads the same runs against each model's own rank from $\tau = 0$ to $\tau = 0.5$: the three models that relax independent censoring move toward rank 1 on all three metrics, the five that assume it move away on oracle IBS and MAE. These are relative ranks, so they say nothing about whether any model's absolute error rises or falls with dependence. They are merely the semi-synthetic counterpart of the crossover the synthetic study isolates with the matched $d_z = 0$ control (Figure 3).



(a) By source dataset. (b) By dependence scenario.

Figure 5: Joint event–censoring recovery. Each point is DVFM's oracle joint-survival ISE divided by HACSurv's, against the true surface $S(e, c \mid \boldsymbol{X})$. Both models share the dataset and split, so they are paired within each of the 60 dataset $\times$ scenario cells after averaging seeds. A ratio below 1 favors DVFM, 0.5 means half the error, and the axis is log-spaced. Dots are geometric means of the ratio; the bar spans the five scenario-specific values in (a) and is a 95% interval across the 12 source datasets in (b), where the label counts the datasets favoring DVFM.

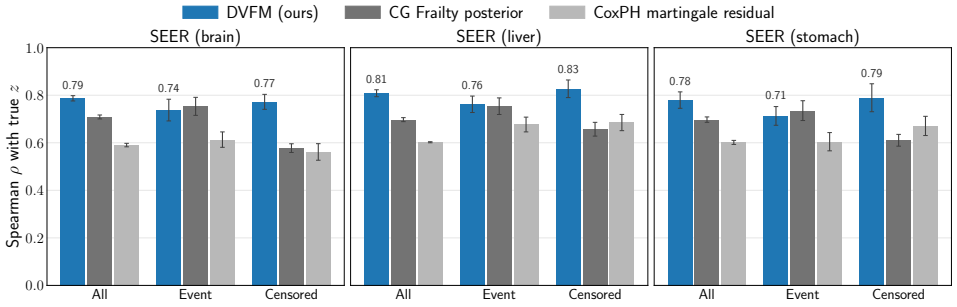


Figure 6: Subject-level recovery of the generating latent against scalar frailty baselines on held-out SEER subjects ($\tau = 0.5$): Spearman correlation with the true z overall and split by censoring status. Bars are means over the four copulas of seed-averaged values, and error bars ± 1 SD across copulas.

Results – Joint Recovery. Figure 5 compares DVFM and HACSurv on the joint surface $S(e, c \mid \boldsymbol{X})$ using paired ISE ratios from identical datasets and splits. DVFM has lower ISE in 37 of 60 dataset $\times$ copula cells. Its mean ISE ratio is 0.91 (95% CI 0.82–1.02; paired Wilcoxon $p = 0.12$), so the difference is not statistically significant. Results vary more across datasets than across copula families: the mean ratio ranges from 0.36 on GBSG to 1.77 on SEER (liver), while the copula-specific means range from 0.85 to 0.97.

Results – Subject-level recovery. Figure 6 compares recovery of the subject-level latent by DVFM, the Cox-Gamma posterior, and the CoxPH martingale residual[2]. Among censored subjects, whose censoring times the baselines treat as uninformative, DVFM's Spearman correlation with the true z is 0.77-0.83 across the three SEER cohorts, against 0.58-0.66 for Cox-Gamma; after an observed event, the two are comparable. Both models learn from event-free survival up to the observed time, but

[2]Both baselines are refit on the same datasets, splits and seeds and scored on the same held-out subjects. For a Cox model with multiplicative $Z \sim \mathrm{Gamma}(\alpha, \alpha)$ frailty (shape–rate), we score the posterior mean log frailty, $\psi(\alpha + \delta_i) - \log[\alpha + r(\boldsymbol{x}_i)\Lambda_0(t_i)]$. For CoxPH, we score the martingale residual, $\delta_i - \widehat{\Lambda}(t_i \mid \boldsymbol{x}_i)$. Within each censoring stratum, both scores decrease with their respective fitted cumulative event hazards.

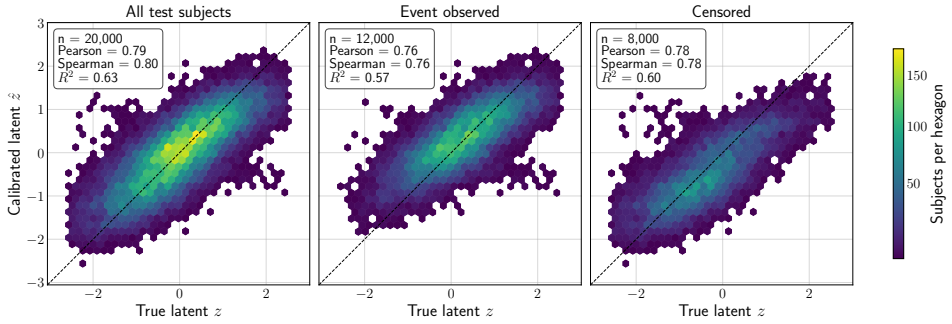


Figure 7: Subject-level recovery of the generating latent on held-out SEER (brain) subjects (Gaussian copula, $\tau = 0.5$, pooled over ten repeats): all subjects (left) and split by censoring status (centre, right). The dashed line marks perfect recovery. Each repeat is calibrated on its own validation split.

only DVFM also uses the censoring process to infer the latent. Figure 7 (SEER brain) and Figure 1 (SUPPORT) show per-subject recovery under a Gaussian copula, and Appendix H.2 across copulas.

Results – Dependence calibration. Estimated τ rises with target dependence but has a dataset-specific offset that grows with the number of uncensored events: on SUPPORT, $\hat{\tau}$ increases from 0.34 at $\tau = 0$ to 0.47 at $\tau = 0.5$. DVFM tracks the direction of change but not the level (Appendix H.3).

7 Real-data experiment

Setup. Finally, we fit DVFM (d_z=1) and the matched d_z=0 ablation to the PRO-ACT dataset to predict time to death in amyotrophic lateral sclerosis (ALS), over 10 seeds with shared splits and hyperparameters. We report the held-out log-likelihood, exact for d_z=0 and a lower bound for d_z=1, with administrative censoring at five years; the margin rows score each margin as if censoring were independent. See Appendices G.3, G.5 and G.7 for the data, metric and training details.

Table 2: Held-out log-likelihood per patient on PRO-ACT (higher is better; mean ± SD over 10 seeds). d_z=1 values are lower bounds.

Log-likelihood	d_z=0	d_z=1	Δ
Joint (t, δ)	-7.03 ± 0.15	$\mathbf{-6.73} \pm 0.03$	$+0.29$
Split by event			
Death	-4.27 ± 0.04	$\mathbf{-4.21} \pm 0.02$	$+0.06$
Censoring	-2.76 ± 0.14	$\mathbf{-2.56} \pm 0.02$	$+0.20$

Results. Table 2 shows the experiment's held-out log-likelihood for DVFM and the no-latent ablation on the PRO-ACT dataset. The latent model scores higher in all 10 splits, with a mean gain of 0.29 per patient. Most of the gain comes from the censoring margin. Since right-censored data alone cannot separate dependent censoring from a better-fitting independent model (Section 4), this shows that the latent improves the fit to observed follow-up, not that censoring in PRO-ACT is dependent.

8 Limitations

DVFM assumes $E \perp\!\!\!\perp C \mid (\boldsymbol{X}, Z)$ and the identification result applies only to a restricted frailty subclass that need not contain the implemented decoder (Theorem 1; Appendix E). Estimated Kendall's τ rises with true dependence but has dataset-specific offsets, so its magnitude is uncalibrated (Section 6). Finally, the semi-synthetic benchmark holds dependence strength constant across covariates and therefore does not test whether DVFM learns covariate-varying dependence.

9 Conclusion

We introduced DVFM, which models event and censoring times through a shared subject-level latent without prescribing a copula family. In synthetic and semi-synthetic datasets, DVFM recovers frailty and improves prediction under dependent censoring, especially for censored subjects. On a real dataset, the latent improves held-out likelihood in every seed. However, choosing when to use the latent remains an open question, since it can reduce predictive accuracy under independent censoring.

AI USE STATEMENT

We used generative AI (Anthropic's Claude) in this work. All AI-assisted output was reviewed and verified by the authors, under the standing constraints documented in `AGENTS.md` in the supplementary material.

Required disclosure. We used AI assistance to implement the methods described here: the encoder and decoder networks, the censored variational objective, the data-generating processes, and the experimental harness. The authors specified the modelling logic, architecture and objective in every case. All such code was reviewed and validated against the reported experiments. We did not use generative AI to generate the synthetic or semi-synthetic datasets (produced algorithmically by the processes in Appendices G.1 and G.2), to develop the theoretical framework, to formulate or prove the mathematical claims, to propose or refine hypotheses, to design the experimental methodology, or for translation or qualitative data analysis.

Recommended disclosure. We used AI assistance to edit the manuscript for readability and consistency, including an audit of mathematical notation that identified overloaded symbols, inconsistent delimiters and undefined notation. The notation table in Appendix A was drafted with AI assistance and checked against the symbols actually used in the source. We also used AI assistance to identify and summarise related work, to produce scientific figures, and for brainstorming. All cited references were verified against their original sources by the authors.

The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses de-identified, publicly available or access-controlled survival datasets. MIMIC-IV was accessed under the PhysioNet credentialed data use agreement, SEER under its standard research data terms, and PRO-ACT under its terms of use. The PRO-ACT trials were approved by the participating medical centers, all patients gave informed consent, and the data were donated for research use and anonymized following HIPAA conventions, including removal of dates and trial-center identity; we made no attempt to re-identify any subject. No identifiable patient information was used and no new data were collected. Predictive results are obtained on simulated or semi-synthetic outcomes, except in the real-data experiment, which uses observed outcomes from completed ALS trials; none of the results are validated for clinical use.

REPRODUCIBILITY STATEMENT

All experiments are specified by the configuration files and run scripts in the supplementary material. The synthetic and semi-synthetic data-generating processes are given in Appendix G.1 and Appendix G.2, with the semi-synthetic construction stated as Algorithm 1; dataset sources and preprocessing are in Appendix G.3. Search spaces and the selection protocol are in Appendix G.8, and metric definitions in Appendix G.5. Every result averages 10 seeds, with seed i passed independently to generation, splitting and model initialization, and all models within a seed share the same dataset and split. Assumptions and full proofs for Theorem 1 are in Appendix E. Experiments used Python 3.11.15 and PyTorch 2.13.0 with CUDA 13.0. DVFM and the neural baselines compute in single precision (fp32); the HACSurv-2D and Cox-Gamma frailty baselines compute in double precision (fp64), as recorded in their configuration defaults. The complete pinned environment is provided in the supplementary material as `environment.lock.yml`, verified to rebuild from scratch.

Repository-level guidance for reproducing the experiments is documented in `README.md` in the supplementary material, with the structural invariants any extension must preserve in `AGENTS.md`.

REFERENCES

Jaap H. Abbring and Gerard J. Van Den Berg. The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):701–710, 2003.

Patricia A. Apellániz, Juan Parras, and Santiago Zazo. Leveraging the variational Bayes autoencoder for survival analysis. *Scientific Reports*, 14:24567, 2024.

Nazem Atassi, James Berry, Amy Shui, Neta Zach, Abby Sherman, Elena Sinani, James Walker, Igor Katsovskiy, David Schoenfeld, and Merit Cudkowicz. The pro-act database: Design, initial analyses, and predictive features. *Neurology*, 83(19):1719–1725, 2014.

Norman E. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review*, pp. 45–57, 1975.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1509.00519.

David G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Carlos Caldas, Samuel Aparicio, and METABRIC Group. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

Angela Dispenzieri, Jerry A. Katzmann, Robert A. Kyle, Dirk R. Larson, Terry M. Therneau, Colin L. Colby, Raynell J. Clark, Graham P. Mead, Shaji Kumar, L. Joseph Melton, and S. Vincent Rajkumar. Use of Nonclonal Serum Immunoglobulin Free Light Chains to Predict Overall Survival in the General Population. *Mayo Clinic Proceedings*, 87(6):517–523, 2012.

Luc Duchateau and Paul Janssen. *The frailty model*. Springer Science & Business Media, 2007.

Takeshi Emura and Yi-Hau Chen. *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Springer, 2018.

Ali H. G. Foomani, Michael Cooper, Russell Greiner, and Rahul G. Krishnan. Copula-based deep survival models for dependent censoring. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216, pp. 669–680, 2023.

Stephane Fotso et al. *PySurvival: Open Source Package for Survival Analysis Modeling*, 2019. Available from https://www.pysurvival.io/.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23):e215–e220, 2000.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, 1999.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85):1–63, 2020.

James J. Heckman and Bo E. Honoré. The identifiability of the competing risks model. *Biometrika*, 76(2):325–330, 1989.

David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis*. John Wiley & Sons, Inc., 2008.

Philip Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396, 1986.

Xuelin Huang and Robert A. Wolfe. A frailty model for informative censoring. *Biometrics*, 58(3): 510–520, 2002.

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.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1–9, 2023.

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*, 18:24, 2018.

David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health. Springer New York, 3 edition, 2012.

William A. Knaus, Frank E. Harrell, Joanne Lynn, Lee Goldman, Russell S. Phillips, Alfred F. Connors, Neal V. Dawson, William J. Fulkerson, Robert M. Califf, Norman Desbiens, Peter Layde, Robert K. Oye, Paul E. Bellamy, Rosemarie B. Hakim, and Douglas P. Wagner. The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults. *Annals of Internal Medicine*, 122(3), 1995.

Christian Marius Lillelund, Ali Hossein Gharari Foomani, Weijie Sun, Shi-ang Qi, Russell Greiner, and The Pooled Resource Open-Access ALS Clinical Trials Consortium (PRO-ACT). MENSA: A Multi-Event Network for Survival Analysis with Trajectory-Based Likelihood Estimation. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297, pp. 543–571, 2026a.

Christian Marius Lillelund, Shi-ang Qi, and Russell Greiner. Overcoming Dependent Censoring in the Evaluation of Survival Models. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337, pp. 3841–3866, 2026b.

Christian Marius Lillelund, Shi-ang Qi, Russell Greiner, and Christian Fischer Pedersen. Position: Stop Chasing the C-index when Evaluating Survival Analysis Models. *arXiv preprint arXiv:2506.02075*, 2026c. Accepted at ICML 2026.

Chuanhui Liu and Xiao Wang. Censor dependent variational inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 39669–39693, 2025.

Xin Liu, Weijia Zhang, and Min-Ling Zhang. HACSurv: A Hierarchical Copula-Based Approach for Survival Analysis with Dependent Competing Risks. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 3079–3087, 2025.

Laura Manduchi, Ričards Marcinkevičs, Michela C. Massi, Thomas Weikert, Alexander Sauter, Verena Gotta, Timothy Müller, Flavio Vasella, Marian C. Neidert, Marc Pfister, Bram Stieltjes, and Julia E Vogt. A Deep Variational Approach to Clustering Survival Data. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RQ428ZptQfU.

National Cancer Institute. Surveillance, epidemiology, and end results (SEER) program. https://seer.cancer.gov, 2024.

David Oakes. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406):487–493, 1989.

Tom Pollard, Benjamin E. Moody, Li-wei Lehman, Brian Gow, Carlos Fernandes, Chenyu Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a Global Platform for Biomedical Research. *Nature Health*, 1:792–795, 2026.

Sebastian Pölsterl. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

Shi-ang Qi, Neeraj Kumar, Mahtab Farrokh, Weijie Sun, Li Hao Kuan, et al. An Effective Meaningful Way to Evaluate Survival Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28244–28276, 2023.

Shi-ang Qi, Yakun Yu, and Russell Greiner. Conformalized Survival Distributions: A Generic Post-Process to Increase Calibration. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 41303–41339, 2024.

Shi-ang Qi, Vahid Balazadeh, Michael Cooper, Russell Greiner, and Rahul G. Krishnan. SurvivalPFN: Amortizing Survival Prediction via In-Context Bayesian Inference. *arXiv preprint arXiv:2605.15488*, 2026.

Patrick Royston and Douglas G. Altman. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13:1–15, 2013.

Terry M. Therneau. *A Package for Survival Analysis in R*, 2024. URL `https://CRAN.R-project.org/package=survival`. R package version 3.6-4.

Anastasios Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22, 1975.

James W. Vaupel, Kenneth G. Manton, and Eric Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454, 1979.

Ruofan Wu, Jiawei Qiao, Mingzhe Wu, Wen Yu, Ming Zheng, Tengfei Liu, Tianyi Zhang, and Weiqiang Wang. Neural frailty machine: beyond proportional hazard assumption in neural survival regressions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5569–5597, 2023.

Zidi Xiu, Chenyang Tao, and Ricardo Henao. Variational learning of individual survival distributions. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL)*, 2020.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. In *Advances in Neural Information Processing Systems*, volume 24, pp. 1845—1853, 2011.

Weijia Zhang, Chun Kai Ling, and Xuanhui Zhang. Deep Copula-Based Survival Analysis for Dependent Censoring with Identifiability Guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20613–20621, 2024.

A NOTATION

Table 3 summarizes all the notation used throughout the paper. Following the convention, we use uppercase letters for random variables and lowercase letters for realizations. Three symbols are reused and disambiguated by subscript or typeface: λ (frailty loading) against λ_E, λ_C (Weibull scales) and Λ_E, Λ_C (cumulative hazards); β (KL weight) against $\boldsymbol{\beta}_E, \boldsymbol{\beta}_C$ (DGP coefficients). We write ϕ only for variational parameters and φ only for a copula generator.

Table 3: Summary of notation.

Symbol	Definition
Data and population quantities	
N, i	Number of subjects; subject index.
d_x, d_z	Covariate dimension; latent dimension.
$\mathcal{D} = \{(\boldsymbol{x}_i, t_i, \delta_i)\}_{i=1}^N$	Observed right-censored dataset.
$\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$, $\boldsymbol{x}_i$	Covariate random vector; realization for subject i.
E, C	Latent event and censoring times, realizations e_i, c_i.
$T = \min(E, C)$	Observed follow-up time, realization t_i.
$\Delta = \mathbb{1}[E \le C]$	Event indicator, realization $\delta_i \in \{0, 1\}$.
$Y = (T, \Delta)$, $y = (t, \delta)$	Observed outcome pair; its realization.
$Z \in \mathbb{R}^{d_z}$, z	Shared latent frailty; its realization.
Model and neural parameterization	
θ, ϕ	Decoder (generative) and encoder (variational) parameters.
p_θ, q_ϕ, $p(z)$	Generative density; variational density; prior $\mathcal{N}(0, I)$.
μ_ϕ, σ_ϕ, ε	Encoder mean, standard deviation, reparameterization noise.
$\psi_E(\boldsymbol{x}, z)$, $\psi_C(\boldsymbol{x}, z)$	Decoder outputs: distributional parameters of each margin.
k_E, k_C; λ_E, λ_C	Weibull shapes; Weibull *scales* (not hazard rates).
Survival and hazard functions	
$S_E(e \mid \boldsymbol{x})$, $S_C(c \mid \boldsymbol{x})$	Marginal conditional survival functions.
$S(e, c \mid \boldsymbol{x})$	Joint conditional survival; unsubscripted S is always the joint.
f_E, f_C	Conditional densities of the two margins.
$\Lambda_E(\cdot, \boldsymbol{x})$, $\Lambda_C(\cdot, \boldsymbol{x})$	Conditional cumulative hazards.
$h(t \mid \cdot)$, $h_0(t)$	Hazard and baseline hazard.
Identification (Appendix E)	
$\eta(z)$, $\zeta(z)$	Frailty effects on the event and censoring margins.
$U = \eta(Z)$, $V = \zeta(Z)$	Realized frailty effects.
H, L_H	Law of (U, V) on $(0, \infty)^2$; its bivariate Laplace transform.
Objective and training	
$\mathcal{L}_1$, $\mathcal{L}_\beta$	Ordinary censored ELBO; β-regularized objective.
β, $\beta_{\max}$, α	KL weight; warm-up ceiling; ℓ_1 shrinkage on latent loadings.
D_{KL}	Kullback-Leibler divergence.
ℓ_r, G_r, K_r	Expected log-likelihood, posterior gap, expected KL of candidate r.
Data-generating process (Appendix G.1)	
$\boldsymbol{\beta}_E$, $\boldsymbol{\beta}_C$; λ	DGP regression coefficients; shared-frailty loading.
$u_{E,i}$, $u_{C,i}$	Independent uniforms used to invert the margins.
c_i^0, a	Uncalibrated censoring draw; censoring-rate calibration constant.
Evaluation (Appendix G.5)	
$\widehat{S}(e, c \mid \boldsymbol{x})$, $\widehat{S}_E(t \mid \boldsymbol{x})$	Estimated joint and marginal event-time survival.
$\widehat{e}_i$, $\widehat{z}_i$	Predicted median event time; posterior-mean latent.
$\mathcal{G}$, $t_{\max}$, t^*	Shared evaluation grid; its endpoint; a time point on it.
τ, $\tau^\star$, $\widehat{\tau}(\boldsymbol{x})$	Kendall's τ: generic, target, model estimate at $\boldsymbol{x}$.
φ	Archimedean copula generator of a fitted baseline.
ISE, BS, IBS, MAE	Integrated squared error; Brier score; integrated Brier score; MAE.
$\mathcal{C}$, $\mathcal{C}_m$	Set of dataset-copula-seed cells; candidate set for model m.

B CENSORING ASSUMPTIONS

Random censoring. Instances censored at time t are representative of all instances still at risk at t, in terms of their survival experience (Kleinbaum & Klein, 2012, Ch. 1). In a medical study, we would

assume that the death rate of censored individuals is the same as that of uncensored individuals who remain in the risk set. Formally, $E \perp\!\!\!\perp C$, where E is the event time and C is the censoring time, meaning $\Pr(E > t \mid C = t) = \Pr(E > t)$ for all t.

Independent Censoring. In the absence of competing risks (*i.e.*, when no alternative event precludes the main event), within any subgroup of interest (*e.g.*, a smoker and a non-smoker group), instances censored at time t are representative of those who remain at risk at time t with respect to their survival experience (Kleinbaum & Klein, 2012). This is also known as conditional independent censoring, since we assume that the observed covariates account for the relationship between event and censoring. Thus, once we condition on the covariates, the censoring time is assumed to be independent of the event time. Formally, $E \perp\!\!\!\perp C \mid \boldsymbol{X}$, where $\boldsymbol{X}$ are the covariates, meaning $\Pr(E > t \mid C = t, \boldsymbol{X}) = \Pr(E > t \mid \boldsymbol{X})$ for all t. Independent censoring is a less restrictive assumption than random censoring on the data-generating process.

Dependent Censoring. In the absence of competing risks, within any subgroup of interest, instances censored at time t are *not* representative of those who remain at risk at time t with respect to their survival experience (Emura & Chen, 2018). Equivalently, censoring is said to be dependent when, even after conditioning on the observed covariates, the censoring mechanism remains related to the event process. Thus, the observed covariates do not fully account for the relationship between E and C, and the censoring time is not independent of the event time given the covariates.

C Derivation of the Censored ELBO

We derive the ELBO for right-censored observations $(t, \delta, \boldsymbol{x})$. Under the generative model

$$p_\theta(e, c, z \mid \boldsymbol{x}) \;=\; p_\theta(e \mid \boldsymbol{x}, z)\, p_\theta(c \mid \boldsymbol{x}, z)\, p(z), \tag{14}$$

we observe (t, δ) where $t = \min(e, c)$ and $\delta = \mathbb{1}(e \le c)$.

Observed likelihood given z. If $\delta = 1$, the event occurs at time t and censoring exceeds t:

$$p(t, \delta = 1 \mid \boldsymbol{x}, z) = f_E(t \mid \boldsymbol{x}, z) S_C(t \mid \boldsymbol{x}, z). \tag{15}$$

If $\delta = 0$, censoring occurs at time t and the event exceeds t:

$$p(t, \delta = 0 \mid \boldsymbol{x}, z) \;=\; S_E(t \mid \boldsymbol{x}, z)\; f_C(t \mid \boldsymbol{x}, z). \tag{16}$$

Combining both cases yields (7).

ELBO. For any variational posterior $q_\phi(z \mid t, \delta, \boldsymbol{x})$,

$$\log p_\theta(t, \delta \mid \boldsymbol{x}) = \log \int p_\theta(t, \delta, z \mid \boldsymbol{x})\, dz \tag{17}$$

$$= \log \int q_\phi(z \mid t, \delta, \boldsymbol{x}) \frac{p_\theta(t, \delta, z \mid \boldsymbol{x})}{q_\phi(z \mid t, \delta, \boldsymbol{x})}\, dz \tag{18}$$

$$\ge \mathbb{E}_{q_\phi}[\log p_\theta(t, \delta, z \mid \boldsymbol{x}) - \log q_\phi(z \mid t, \delta, \boldsymbol{x})] \tag{19}$$

$$= \mathbb{E}_{q_\phi}[\log p_\theta(t, \delta \mid \boldsymbol{x}, z)] - D_{\mathrm{KL}}(q_\phi(z \mid t, \delta, \boldsymbol{x}) \| p(z))\,. \tag{20}$$

This derivation yields the ordinary censored-data ELBO $\mathcal{L}_1$ in (8). The training criterion $\mathcal{L}_\beta$ in (10) is a regularized variant. When $\beta \neq 1$, it is not in general an evidence lower bound; free-bits modifications change the criterion further and are treated as optimization regularizers.

D Proof of Proposition 1

Proof. Conditional independence gives

$$S(e, c \mid \boldsymbol{x}) = \mathbb{E}_Z\left[S_E(e \mid \boldsymbol{x}, Z) S_C(c \mid \boldsymbol{x}, Z)\right].$$

The marginal survival functions satisfy

$$S_E(e \mid \boldsymbol{x}) = \mathbb{E}_Z[S_E(e \mid \boldsymbol{x}, Z)], \qquad S_C(c \mid \boldsymbol{x}) = \mathbb{E}_Z[S_C(c \mid \boldsymbol{x}, Z)].$$

Subtracting their product proves the covariance identity. Independence is equivalent to factorization of the joint survival function for every pair of thresholds. □

E PROOF OF THEOREM 1

This appendix states the assumptions and theorem summarized in Section 4, then gives the proof in full, followed by two remarks on the variation condition and on the scope of the result.

Assumption 1 (Conditional independence and multiplicative frailty)**.** *There exist measurable* frailty-effect *functions η and ζ such that $U = \eta(Z)$ and $V = \zeta(Z)$ are finite and strictly positive almost surely. Conditional on (X, Z), the event and censoring times are independent, and*

$$S_E(\, e \,|\, \boldsymbol{x}, z) = \exp\{-\Lambda_E(e, \boldsymbol{x})\, \eta(z)\}, \tag{21}$$

$$S_C(\, c \,|\, \boldsymbol{x}, z) = \exp\{-\Lambda_C(c, \boldsymbol{x})\, \zeta(z)\}. \tag{22}$$

For $j \in \{E, C\}$, the function Λ_j is finite and jointly continuous on $[0, \infty) \times \mathcal{X}$. For each $\boldsymbol{x}$, it is nonnegative and nondecreasing in time, satisfies $\Lambda_j(0, \boldsymbol{x}) = 0$, and tends to infinity as time tends to infinity.

Assumption 2 (Covariate-invariant frailty-effect distribution)**.** *The conditional distribution of (U, V) given $\boldsymbol{X} = \boldsymbol{x}$ is a single, possibly unknown, probability measure H on $(0, \infty)^2$, common to all $\boldsymbol{x} \in \mathcal{X}$. Thus, for every Borel set $A \subseteq (0, \infty)^2$,*

$$\Pr\big((U, V) \in A \mid \boldsymbol{X} = \boldsymbol{x}\big) = H(A).$$

In particular, $(U, V) \perp\!\!\!\perp \boldsymbol{X}$.

Assumption 3 (Known indices and two-dimensional support variation)**.** *The functions Λ_E and Λ_C, including their scales, are known and are held fixed across candidate models. The set*

$$\mathcal{A} = \big\{\big(\Lambda_E(t, \boldsymbol{x}), \Lambda_C(t, \boldsymbol{x})\big) : t > 0,\ \boldsymbol{x} \in \mathcal{X}\big\}$$

contains a nonempty open set $\mathcal{O}$ in the strictly positive quadrant $(0, \infty)^2$.

Theorem 1 (Identification conditional on known hazard indices)**.** *Under Assumptions 1- 3, the population distribution of $(T, \boldsymbol{X})$ uniquely identifies H. Consequently, for every $\boldsymbol{x} \in \mathcal{X}$ and $e, c \geq 0$, it identifies*

$$S(e, c \mid \boldsymbol{x}) = L_H\big(\Lambda_E(e, \boldsymbol{x}), \Lambda_C(c, \boldsymbol{x})\big), \tag{23}$$

where

$$L_H(a, b) = \int_{(0,\infty)^2} \exp(-au - bv)\, H(du, dv), \qquad a, b \geq 0. \tag{24}$$

The event and censoring marginal survival functions are identified as

$$S_E(e \mid \boldsymbol{x}) = L_H\big(\Lambda_E(e, \boldsymbol{x}), 0\big), \tag{25}$$

$$S_C(c \mid \boldsymbol{x}) = L_H\big(0, \Lambda_C(c, \boldsymbol{x})\big). \tag{26}$$

Proof. Conditional independence and the covariate-invariance of H give

$$\begin{aligned} S_{\text{obs}}(t \mid \boldsymbol{x}) &= \Pr(T > t \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \Pr(E > t, C > t \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \int_{(0,\infty)^2} \exp\{-\Lambda_E(t, \boldsymbol{x})u - \Lambda_C(t, \boldsymbol{x})v\}\, H(du, dv) \\ &= L_H\big(\Lambda_E(t, \boldsymbol{x}), \Lambda_C(t, \boldsymbol{x})\big). \end{aligned} \tag{27}$$

Consider two candidate laws H_1 and H_2 satisfying the assumptions with the same known hazard indices and inducing the same distribution of $(T, \boldsymbol{X})$. Their conditional observed survival functions coincide for $P_{\boldsymbol{X}}$-almost every $\boldsymbol{x}$. Joint continuity of the indices and dominated convergence make the model-implied survival functions continuous in $(t, \boldsymbol{x})$. Equality therefore extends to all $\boldsymbol{x} \in \mathcal{X}$ and all $t \geq 0$: first take a common full-measure set for rational t, then use continuity in $\boldsymbol{x}$ on the support and continuity in t. Consequently,

$$L_{H_1}(a, b) = L_{H_2}(a, b) \qquad \text{for every } (a, b) \in \mathcal{O}.$$

Laplace transforms of finite measures on $(0, \infty)^2$ are real analytic on $(0, \infty)^2$. This follows from expanding the exponential locally: around (a_0, b_0), increments satisfying $|h| < a_0/2$ and $|k| < b_0/2$ give an absolutely integrable power series, since

$$\exp(-a_0 u - b_0 v) \exp(|h|u + |k|v) \leq 1.$$

No unweighted moment assumptions on H_1 or H_2 are needed. Choose nonempty open intervals I, J with $I \times J \subseteq \mathcal{O}$. For each $b \in J$, the one-dimensional identity theorem extends equality in a from I to $(0, \infty)$. For each resulting $a > 0$, applying the same theorem in b extends equality from J to $(0, \infty)$. Hence

$$L_{H_1}(a, b) = L_{H_2}(a, b) \qquad \text{for all } a, b > 0.$$

Continuity extends this equality to $a, b \geq 0$.

For completeness, uniqueness of the mixing law can be seen by setting $R = \exp(-U)$ and $W = \exp(-V)$. Equality of the transforms implies equality of all mixed integer moments $\mathbb{E}[R^m W^n]$, $m, n \geq 0$. Polynomials are dense in $C([0, 1]^2)$, so the induced probability measures of (R, W) coincide. Since $(u, v) \mapsto (\exp(-u), \exp(-v))$ is one-to-one with a measurable inverse on its image, $H_1 = H_2$.

Finally, conditional independence gives (23). The identified H and known indices therefore determine the joint survival function and both marginal survival functions. □

Scope. The identified object is the law of the frailty effects $(\eta(Z), \zeta(Z))$, not a latent coordinate system, the maps η, ζ separately, or individual realized frailties; fixing $Z \sim \mathcal{N}(0, I)$ does not make H known. The argument follows Abbring & Van Den Berg (2003, Prop. 2(b)), who relax the full-quadrant condition of Heckman & Honoré (1989) to an open set by the same analyticity argument and do not assume the indices known; they require proportional indices $\lambda(t)\phi(\boldsymbol{x})$ and finite frailty means. Assuming the indices known lets them be non-separable in $(t, \boldsymbol{x})$ and removes the moment condition, but is otherwise a specialization of their result. The result is a sufficient condition for a restricted subclass: the implemented decoder, whose Weibull shape depends on z, need not satisfy the multiplicative separation of Assumption 1, and population uniqueness implies neither finite-sample accuracy nor convergence of variational training.

Remark 1 (Meaning of the variation condition)**.** *Assumption 3 concerns the population covariate support, not the finite set of covariates in a sample. For example, with a scalar covariate whose support contains an open interval I_x, the known indices*

$$\Lambda_E(t, x) = t, \qquad \Lambda_C(t, x) = t \exp(x)$$

produce the open image $\{(a, b) : a > 0,\ \log(b/a) \in I_x\}$. By contrast, if $\Lambda_C(t, x) = k\Lambda_E(t, x)$ for a fixed $k > 0$, the observable transform values lie on a single ray:

$$L_H(a, ka) = \mathbb{E}[\exp\{-a(U + kV)\}].$$

These values identify at most the law of $U + kV$, not the joint law of (U, V) in general. The open-set condition is sufficient for the present argument; it is not asserted to be necessary for every more restricted model. Heckman & Honoré (1989) require the covariate index to cover the full quadrant $(0, \infty)^2$; Abbring & Van Den Berg (2003) weaken this to the same open-set requirement used here, in the mixed proportional hazards model with unknown indices.

Remark 2 (Scope for DVFM)**.** *The identified object is the distribution of the frailty effects $(\eta(Z), \zeta(Z))$, not a unique latent coordinate system, the functions η and ζ separately, individual realized frailties, or neural-network parameters. Fixing a prior such as $Z \sim \mathcal{N}(0, I)$ does not make H known when the mappings η and ζ are unknown.*

The theorem does not establish identification of unknown hazard indices jointly with H. It remains applicable after a separately proved identification of those indices, with their scales fixed, but merely estimating them with a neural network does not supply that proof. Likewise, allowing the frailty-effect distribution to vary freely with $\boldsymbol{x}$ removes the common transform used in the argument.

F Objective-Level Theory for DVFM

Let $Y = (T, \Delta)$, with realization $y = (t, \delta)$, and define the ordinary censored-data ELBO

$$\mathcal{L}_1(\theta, \phi; y, \boldsymbol{x}) = \mathbb{E}_{q_\phi(z \mid y, \boldsymbol{x})}[\log p_\theta(y \mid \boldsymbol{x}, z)] - D_{\mathrm{KL}}(q_\phi(z \mid y, \boldsymbol{x}) \| p(z)) \,. \tag{28}$$

The identity

$$\mathcal{L}_1(\theta, \phi; y, \boldsymbol{x}) = \log p_\theta(y \mid \boldsymbol{x}) - D_{\mathrm{KL}}(q_\phi(z \mid y, \boldsymbol{x}) \| p_\theta(z \mid y, \boldsymbol{x})) \tag{29}$$

shows that $\mathcal{L}_1$ is a lower bound on the observed-data log likelihood. If the KL term is multiplied by $\beta \neq 1$, then

$$\mathcal{L}_\beta = \mathcal{L}_1 + (1-\beta) D_{\mathrm{KL}}(q_\phi(z \mid y, \boldsymbol{x}) \| p(z)) \,, \tag{30}$$

so $\mathcal{L}_\beta$ is a regularized variational objective rather than an ordinary ELBO.

Proposition 2 (Existence, but not uniqueness, of a latent-inactive optimum)**.** *Assume $E \perp\!\!\!\perp C \mid \boldsymbol{X}$ and that the true observed-data density $p_0(y \mid \boldsymbol{x})$ belongs to the latent-inactive submodel: there exists θ_0 such that*

$$p_{\theta_0}(y \mid \boldsymbol{x}, z) = p_0(y \mid \boldsymbol{x})$$

for $p(z)$-almost every z. Assume also that the variational family contains $q_\phi(z \mid y, \boldsymbol{x}) = p(z)$, and write ϕ_0 for such a ϕ. Then, for $\beta = 1$, the population ELBO is maximized at (θ_0, ϕ_0) and its value is

$$\mathbb{E}_{p_0} \log p_0(Y \mid \boldsymbol{X}).$$

This establishes the existence of a zero-dependence optimum. It does not establish uniqueness and does not imply that every population or finite-sample optimizer has zero model-implied dependence.

Proof. For every (θ, ϕ),

$$\begin{aligned}
\mathbb{E}_{p_0} \mathcal{L}_1(\theta, \phi; Y, \boldsymbol{X}) &= \mathbb{E}_{p_0} \log p_\theta(Y \mid \boldsymbol{X}) - \mathbb{E}_{p_0} D_{\mathrm{KL}}(q_\phi(z \mid Y, \boldsymbol{X}) \| p_\theta(z \mid Y, \boldsymbol{X})) && (31)\\
&\leq \mathbb{E}_{p_0} \log p_\theta(Y \mid \boldsymbol{X}) && (32)\\
&\leq \mathbb{E}_{p_0} \log p_0(Y \mid \boldsymbol{X}), && (33)
\end{aligned}$$

where the last inequality is the conditional cross-entropy inequality. At the latent-inactive parameter θ_0, the exact posterior equals the prior, so choosing $q_{\phi_0} = p$ makes both inequalities equalities. Different latent-variable decompositions can nevertheless induce the same observed density and attain the same value; right-censored data alone do not select among such observationally equivalent decompositions. □

Proposition 3 (Comparison of frailty and proxy explanations)**.** *For a candidate $r \in \{\mathrm{fr}, \mathrm{pr}\}$, define*

$$\begin{aligned}
\ell_r &= \mathbb{E}_{p_0} \log p_{\theta_r}(Y \mid \boldsymbol{X}), && (34)\\
G_r &= \mathbb{E}_{p_0} D_{\mathrm{KL}}(q_{\phi_r}(z \mid Y, \boldsymbol{X}) \| p_{\theta_r}(z \mid Y, \boldsymbol{X})) \,, && (35)\\
K_r &= \mathbb{E}_{p_0} D_{\mathrm{KL}}(q_{\phi_r}(z \mid Y, \boldsymbol{X}) \| p(z)) \,. && (36)
\end{aligned}$$

Then

$$\mathbb{E}_{p_0} \mathcal{L}_\beta(\theta_{\mathrm{fr}}, \phi_{\mathrm{fr}}) - \mathbb{E}_{p_0} \mathcal{L}_\beta(\theta_{\mathrm{pr}}, \phi_{\mathrm{pr}}) = (\ell_{\mathrm{fr}} - \ell_{\mathrm{pr}}) - (G_{\mathrm{fr}} - G_{\mathrm{pr}}) + (1-\beta)(K_{\mathrm{fr}} - K_{\mathrm{pr}}). \tag{37}$$

Consequently, at $\beta = 1$, the frailty candidate is preferred precisely when its observed-data log-likelihood advantage exceeds its excess variational gap. If the two candidates induce the same observed-data distribution and both posteriors are represented exactly, the population ELBO ties them.

Proof. Substitute $\mathbb{E}_{p_0} \mathcal{L}_1 = \ell - G$ and $\mathcal{L}_\beta = \mathcal{L}_1 + (1-\beta)K$, then subtract the two candidate objectives. □

Interpretation. Proposition 3 is a comparison result, not a dependence-recovery theorem. The reconstruction term by itself is not a marginal cross-entropy: an encoder can use (T, Δ) to reduce reconstruction loss even when censoring is independent. Any stronger recovery claim therefore requires an explicit structural restriction, a sensitivity analysis, or information beyond ordinary right-censored observations.

G Experimental details

G.1 Synthetic experiments

The synthetic study generates both times from one known shared Gaussian frailty, so the latent DVFM infers can be compared against the latent that produced the data and the true event time is available for every subject.

Generative process. Coefficients $\boldsymbol{\beta}_E, \boldsymbol{\beta}_C \sim \mathcal{N}(0, 0.25^2 I_{d_x})$ are drawn once per DGP seed and held fixed across dependence and censoring levels. We then draw $N = 10{,}000$ covariate vectors $\boldsymbol{x}_i \sim \mathcal{N}(0, I_{d_x})$ with $d_x = 10$, one frailty $z_i \sim \mathcal{N}(0,1)$ per subject, and independent $u_{E,i}, u_{C,i} \sim \mathrm{Unif}(0,1)$. Event and baseline censoring times take the Weibull accelerated-failure-time form

$$e_i = \exp\big(\boldsymbol{x}_i^\top \boldsymbol{\beta}_E + \lambda z_i\big)(-\log u_{E,i})^{1/k_E}, \qquad c_i^0 = \exp\big(\boldsymbol{x}_i^\top \boldsymbol{\beta}_C + \lambda z_i\big)(-\log u_{C,i})^{1/k_C}, \quad (38)$$

with shapes $k_E = 1.5$ and $k_C = 1.3$, so conditional on $(\boldsymbol{x}_i, z_i)$ each time is Weibull with that shape and scale $\exp\big(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \lambda z_i\big)$. The single loading λ multiplies the same z_i in both predictors and is the only channel coupling the two times, so (1) holds by construction. Writing $S_E(e \mid \boldsymbol{x}, z) = \exp\{-\Lambda_E(e, \boldsymbol{x})\, \eta(z)\}$ with $\Lambda_E(e, \boldsymbol{x}) = \big(e \exp(-\boldsymbol{x}^\top \boldsymbol{\beta}_E)\big)^{k_E}$ and $\eta(z) = \exp(-k_E \lambda z)$, and likewise for C with $\zeta(z) = \exp(-k_C \lambda z)$, shows the construction satisfies Assumptions 1 and 2 exactly: the frailty effects are lognormal, drawn independently of $\boldsymbol{X}$, and perfectly dependent, since $\zeta(z) = \eta(z)^{k_C/k_E}$. The censoring calibration below rescales c^0 by a constant e^a, which enters $\Lambda_C(c, \boldsymbol{x}) = \big(c \exp(-\boldsymbol{x}^\top \boldsymbol{\beta}_C - a)\big)^{k_C}$ as a scale on the index and leaves the frailty effect unchanged. The joint survival function is therefore $L_H(\Lambda_E(e, \boldsymbol{x}), \Lambda_C(c, \boldsymbol{x}))$ as in (23), with L_H the Laplace transform of that lognormal mixing law. It has no closed form, so the induced copula is a lognormal frailty copula rather than a member of the standard named families.

Dependence calibration. No copula is imposed. Instead λ is calibrated to a target conditional Kendall's τ by bisection on $[0, 16]$ over 24 steps, each step estimating τ from $50{,}000$ draws of the residual pair $\big(\lambda z + k_E^{-1} \log(-\log u_E),\ \lambda z + k_C^{-1} \log(-\log u_C)\big)$, and $\lambda = 0$ recovers $E \perp\!\!\!\perp C \mid \boldsymbol{X}$ exactly. The induced dependence is a normal mixture of independent Gumbel errors rather than a Gaussian copula; we record the achieved conditional and marginal τ for every generated dataset.

Censoring calibration. Writing $c_i = e^a c_i^0$, subject i is censored exactly when $\log e_i - \log c_i^0 > a$, so setting a to the empirical $(1 - \text{rate})$-quantile of that difference attains the target rate to within one subject and leaves the dependence untouched. The observed data are $t_i = \min(e_i, c_i)$ and $\delta_i = \mathbb{1}[e_i \le c_i]$; the true e_i and z_i are retained for oracle evaluation only.

Grid and fitting. We cross $\tau \in \{0, 0.25, 0.5, 0.75\}$ with censoring rates $\{0.25, 0.50, 0.75\}$ over 10 repeats, splitting each dataset $70/10/20$. Each dataset is fitted twice, at $d_z = 1$ and at the matched $d_z = 0$ control (Appendix G.7); the two variants receive identical times, indicators and splits, so every comparison is paired within a repeat.

G.2 Semi-synthetic experiments

We evaluate DVFM on semi-synthetic datasets that pair real covariates with simulated outcomes. The construction serves two purposes at once. It fixes the dependence between the event and censoring processes at a known Kendall's τ under a chosen copula family, so that a model's estimate of dependence can be compared against a target; and it produces a true event time for every subject, including those the sampled censoring time hides, so that oracle metrics are computable on the same footing for all models. Neither is available from a real dataset. What is retained from the real dataset is the covariate matrix, unchanged apart from preprocessing, and the shape of the event and censoring margins, which are fit to it.

Algorithm 1 states the procedure. It follows Foomani et al. (2023, Algorithm 4) in structure – fit margins, draw dependent uniforms from a copula, invert the margins by Sklar's theorem – and departs from it in four places. First, their label is the true event time, so they fit one Weibull model and obtain the censoring model by shrinking its variance; we have no uncensored label, so we fit two Cox margins to the source dataset, the event margin on δ and the censoring margin on $1 - \delta$. Second, we resample both times rather than only the censoring time, which is what makes the generating latent and the true event time available for every subject. Third, we calibrate the censoring rate to a target instead of using their fixed variance heuristic: writing C_0 for the uncalibrated censoring draw, the rescaled time $C = s \cdot C_0$ censors subject i exactly when $E_i / C_{0,i} > s$, so the empirical $(1 - \text{rate})$-quantile of that ratio hits the target to within one subject whenever the ratios are untied, and leaves the copula untouched. Because the inverted Breslow margins place a point mass at the first source time, ties can occur under dependence; the achieved rate then falls below the target by the number of tied subjects, by at most 0.9 percentage points (MIMIC-IV) across all seeds and copulas at $\tau = 0.5$. Fourth, we sample each family through its shared-factor representation and record the shared variable, standardized, as the recovery target z.

DVFM is misspecified throughout this benchmark: the margins are CoxPH fits rather than its Weibull decoder; the mixing variable is Gamma, logarithmic-series, positive-stable or Gaussian rather than its standard normal prior; and dependence enters through a fixed generator on marginal quantiles rather than a neural decoder of $(\boldsymbol{x}, z)$.

Algorithm 1 Semi-synthetic dataset construction: PyTorch-like Pseudocode

```
# d_src: real right-censored dataset (X, T, delta)
# copula: family with Kendall's tau in [0, 1)
# rate: target censoring rate
def make_semi_synthetic(d_src, copula, tau, rate, seed):

    # Impute, z-score numerics, one-hot categoricals; X is kept as-is
    X = preprocess(d_src.X)

    # Two Cox margins fit to the source dataset
    m_e = CoxPH(pen=0.01).fit(X, d_src.T, d_src.delta) # S_E(e|x)
    m_c = CoxPH(pen=0.01).fit(X, d_src.T, 1 - d_src.delta) # S_C(c|x)

    # Shared-factor sampling: w generates the copula
    u, v, w = copula.sample(n=len(X), tau=tau, seed=seed)
    z = standardize(w if copula.is_gaussian else log(w))

    # Sklar's theorem: invert each margin at its own uniform
    E = m_e.inverse_survival(u, X) # solves S_E(E|x) = u
    C0 = m_c.inverse_survival(v, X) # solves S_C(C0|x) = v

    # C = s * C0 censors subject i exactly when E_i / C0_i > s,
    # so one scalar quantile calibrates the rate
    s = quantile(E / C0, 1 - rate)
    C = s * C0

    T, delta = min(E, C), (E <= C)
    return Dataset(X, T, delta, true_E=E, true_C=C, true_z=z)
```

Notes: `inverse_survival` inverts the fitted Breslow cumulative hazard by interpolation, extrapolating linearly beyond the last observed source time. `copula.sample` draws (u, v) from the common normal factor for the Gaussian copula and from the Marshall-Olkin mixing variable for Clayton, Frank and Gumbel; w is that shared variable. `standardize` centers and scales to unit variance. `quantile` is the empirical quantile over subjects.

Figures 8 and 9 compare the original and generated datasets for a representative repeat. We show empirical KM curves and observed follow-up times stratified by event status. Under dependent censoring, the generated KM curves may differ from the event-time distribution because the ordinary KM estimator assumes independent censoring. Reported censoring rates are computed from the generated event indicators. Because the two margins are fit independently of the copula, these figures check the marginal fit only; the copula governs the association between the two times, not either margin, so agreement here is not evidence that the dependence was generated correctly.

G.3 Datasets and preprocessing

Table 4 summarizes the 12 semi-synthetic source datasets and the PRO-ACT dataset of the real-data experiment; the paragraphs below give provenance and preprocessing for each. To keep computation manageable, we subsample the largest datasets before fitting the semi-synthetic event and censoring models. For Employee, MIMIC-IV, and each SEER cohort, we first remove records with non-positive or non-finite follow-up times and then retain exactly 10,000 individuals. To ensure that the subsample remains representative, we divide follow-up times into 10 quantile-based groups and form strata from the combination of follow-up-time group and event status. We then randomly sample from each stratum in proportion to its size in the original dataset, without selecting any individual more than once. A deterministic rounding procedure ensures that the final subsample contains exactly 10,000 individuals. This preserves the original distribution of event status and observed follow-up time as closely as possible.

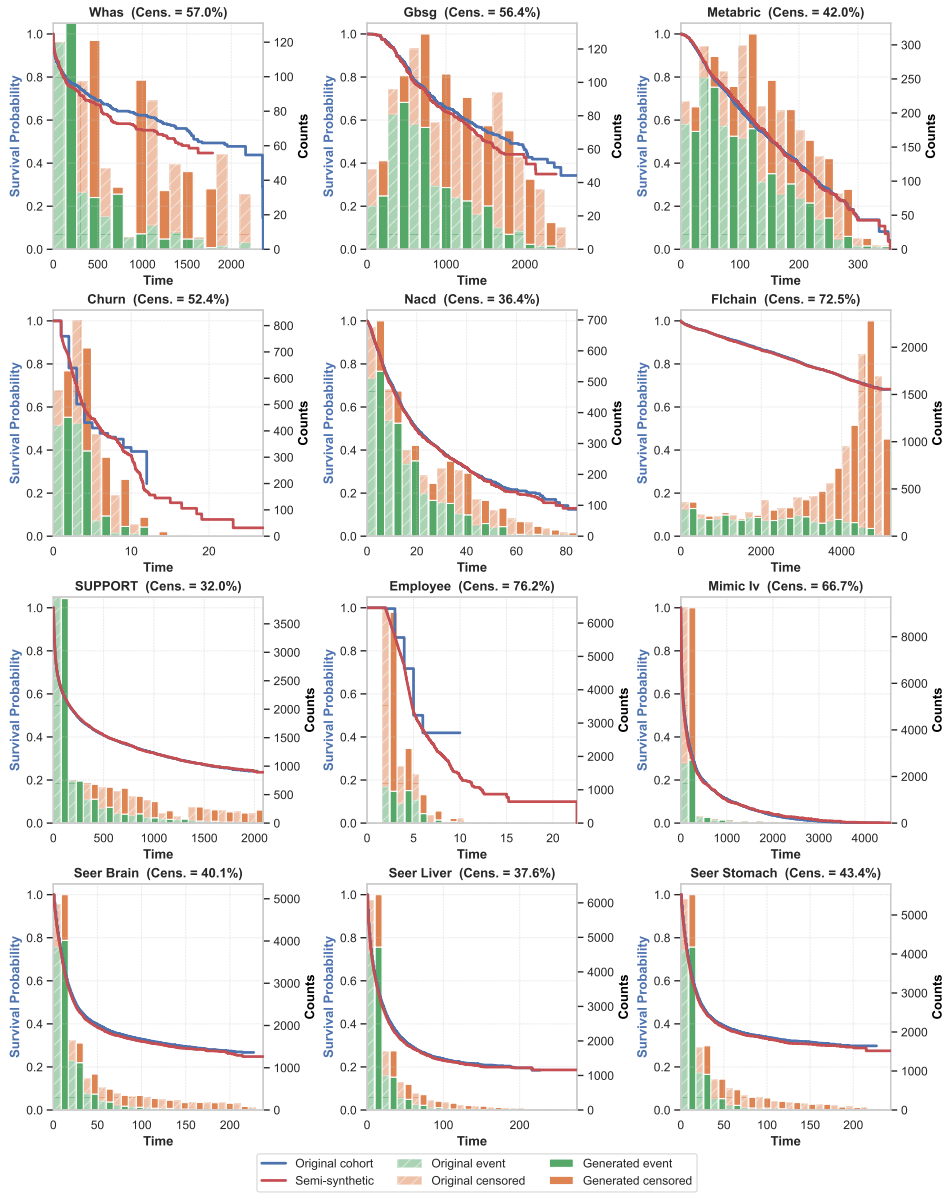


Figure 8: Original and learned event/censoring distributions under independent censoring ($\tau = 0$). Lines show empirical KM curves, and histograms show observed follow-up times stratified by event status.

WHAS: Worcester Heart Attack Study dataset (WHAS) (Hosmer et al., 2008) contains 500 patients with acute myocardial infarction, focusing on the time to death post-hospital admission. The data was already post-processed and can be downloaded from the `scikit-survival` package (Pölsterl, 2020).

METABRIC: The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset (Curtis et al., 2012) contains survival information for breast cancer patients. This dataset includes a diverse range of feature sets that encompass clinical traits, expression profiles, copy number variation (CNV) profiles, and single-nucleotide polymorphism (SNP) genotypes. All these features are derived from breast tumor samples collected during the METABRIC trial. The dataset can be downloaded from `https://www.cbioportal.org/study/summary?id=brca_metabric`.

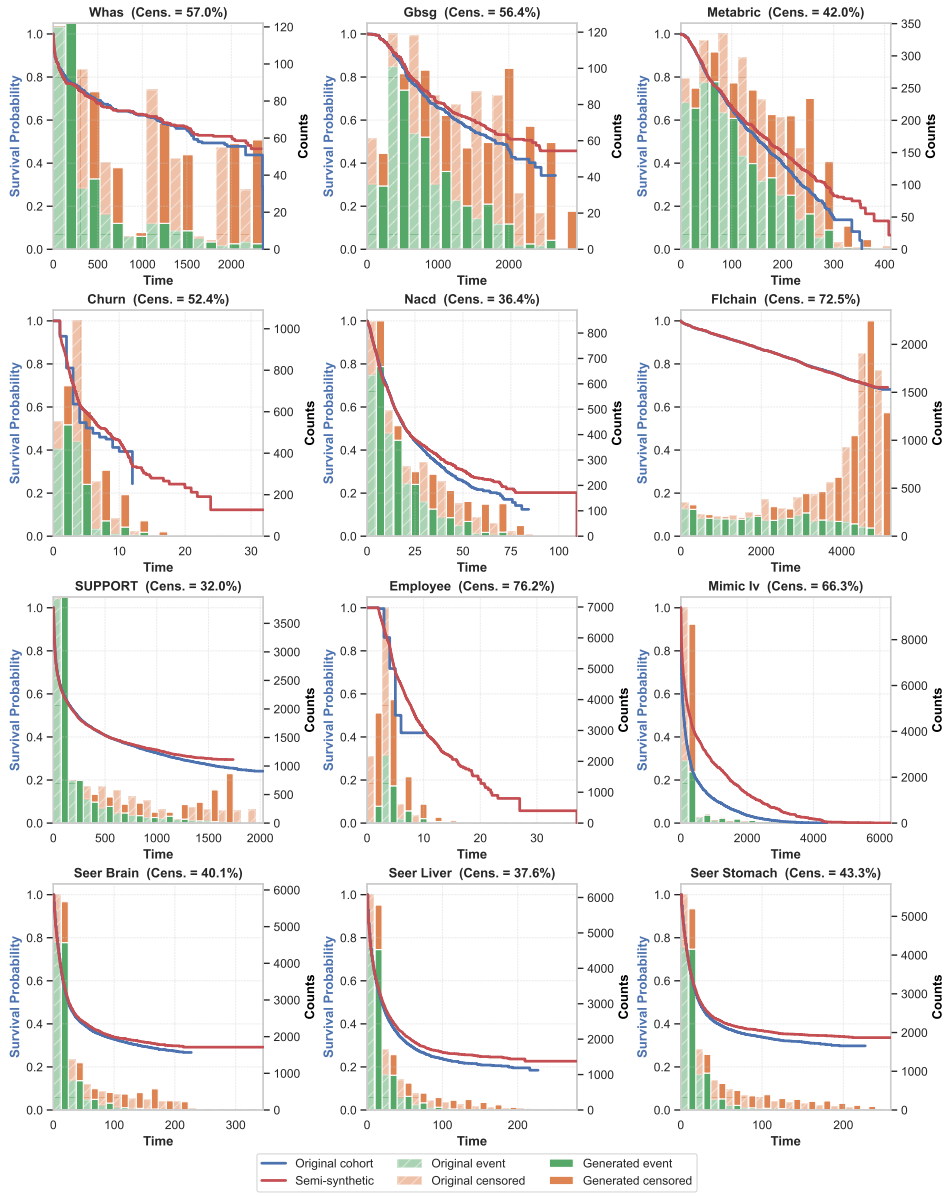


Figure 9: Original and learned event/censoring distribution under dependent censoring ($\tau = 0.5$). Dependent censoring changes the observed-data KM curves even though the generated event-time marginals are unchanged by construction.

Churn: Customer churn prediction dataset (Churn) focuses on predicting customer attrition. Each record in the file corresponds to a customer and contains the same set of features, such as age, gender, tenure, usage frequency, support calls, payment delay, subscription type, contract length, total spend, and last interaction. We apply one-hot encoding on the US region feature and exclude subjects who are censored at time 0. The dataset can be downloaded from the PySurvival package (Fotso et al., 2019).

GBSG: The German Breast Cancer Study Group (GBSG) dataset (Royston & Altman, 2013) includes 686 patients with node-positive breast cancer and complete prognostic information, with recurrence-free survival as the outcome. We use the version distributed in R's `survival` package (Therneau, 2024), with the eight covariates age, menopausal status, tumour size, grade, number of positive

Table 4: Overview of the survival datasets after preprocessing. †Number of features before (after) one-hot encoding. ‡Stratified subsample of exactly 10,000 subjects drawn before the semi-synthetic margins are fit. §Real-data experiment only; outcomes are observed, not generated.

Dataset	N	$p^{\dagger}$	Cens. (%)
WHAS	500	14 (14)	57.0
GBSG	686	8 (8)	56.4
METABRIC	1,902	9 (9)	42.1
Churn	1,958	12 (25)	52.4
NACD	2,396	48 (48)	36.4
PRO-ACT§	3,172	8 (10)	38.1
FLCHAIN	7,871	9 (31)	72.5
SUPPORT	8,873	14 (27)	32.0
Employee‡	14,999	8 (17)	76.2
MIMIC-IV‡	38,520	91 (103)	66.7
SEER-brain‡	73,703	10 (10)	40.1
SEER-liver‡	82,841	14 (14)	37.6
SEER-stomach‡	100,360	14 (14)	43.4

nodes, progesterone and oestrogen receptor levels, and hormonal therapy. No further preprocessing is applied beyond dropping the patient identifier.

NACD: The Northern Alberta Cancer Dataset (NACD) (Haider et al., 2020) contains 2402 patients and 53 features. It is a conglomerate of many different cancer patients, including lung, colorectal, head and neck, esophageal, stomach, and other cancers. The event of interest in this dataset is failure time. We drop patients with negative or zero survival time leaving 2,396 patients. There are no missing values in this dataset. The dataset can be downloaded from `http://pssp.srv.ualberta.ca` under "Public Predictors".

FLCHAIN: The Serum Free Light Chain (FLCHAIN) dataset is a stratified random sample comprising half of the participants from a study investigating the association between serum free light chain (FLC) levels and mortality (Dispenzieri et al., 2012). It is available in R's `survival` package (Therneau, 2024). After downloading, we perform several preprocessing steps: we remove three subjects with events occurring at time zero, impute missing values in the `creatinine` feature using its median, and exclude the `chapter` feature—which encodes the cause of death according to ICD chapter headings—since it is only defined for deceased (uncensored) subjects and would therefore leak the event indicator to the model.

SUPPORT: The Study to Understand Prognoses, Preferences, Outcomes, and Risks of Treatment (SUPPORT) dataset (Knaus et al., 1995) includes 9,105 participants and was designed to investigate survival outcomes and clinical decision-making among seriously ill hospitalized patients. The dataset contains a substantial proportion of missing values across many features. We obtained the dataset from Qi et al. (2024) and follow their preprocessing steps.

Employee: Employee dataset contains employee activity information that can be used to predict when an employee will quit. The dataset can be downloaded from the PySurvival package (Fotso et al., 2019). The file contains 14,999 records, of which 3,008 are exact duplicates; we retain them, as the source distributes them, before drawing the stratified 10,000-subject subsample described above.

MIMIC-IV: The Medical Information Mart for Intensive Care IV (MIMIC-IV) (Johnson et al., 2023), distributed through PhysioNet (Pollard et al., 2026; Goldberger et al., 2000), contains critical care data from patients admitted to hospitals and intensive care units (ICU). We create a version using the MIMIC-IV database, which contains patients who are alive at least 24 hours after being admitted to ICU. Their date of death is derived from hospital records or state records, which means the cause of mortality is not limited to the reason for ICU admission. We follow the instructions in (Qi et al., 2023) to process the dataset, and the code is available in their GitHub repository: `https://github.com/shi-ang/CensoredMAE`.

SEER: We use the SEER-brain, SEER-liver and SEER-stomach datasets of Qi et al. (2024), derived from the Surveillance, Epidemiology, and End Results (SEER) Program (National Cancer Institute,

2024), and follow their preprocessing. The datasets cover patients diagnosed with brain, liver and stomach cancers, with the outcome being time from diagnosis to death. Features include age, sex, cancer stage, grade and other clinical factors; features with over 70% missing data, patients without follow-up times, duplicates, and patients with zero survival time were excluded. The SEER data are available at `https://seer.cancer.gov`.

PRO-ACT: The Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) database (Atassi et al., 2014) pools de-identified records from completed amyotrophic lateral sclerosis trials and is accessed under its terms of use; unlike the other datasets, it is used with its observed outcomes. Data used in the preparation of this article were obtained from the PRO-ACT Database; the data available in the PRO-ACT Database have been volunteered by PRO-ACT Consortium members. The event is death, from the mortality form, and the censoring time is the last day a patient appears on a clinical assessment form; laboratory, adverse-event and medication dates are excluded, since they include entry errors and incompletely recorded records that can postdate every assessment. Of 8,098 patients with an ALSFRS record and a recorded onset, we exclude 4,733 without a recorded vital status, 166 deceased without a death day and 27 without follow-up, leaving 3,172 patients and 1,962 deaths. The eight baseline covariates follow Lillelund et al. (2026a); an onset under 30 days before screening is set to missing. Missingness is largely trial-level (up to 43% per covariate); values are imputed with the training-split mean or mode, and each covariate missing in the training split receives a 0/1 indicator. Each seed uses its own $70/10/20$ split, stratified by event status and observed-time band.

Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database; the data available in the PRO-ACT Database have been volunteered by PRO-ACT Consortium members.

G.4 Baselines learners

CoxPH: The Cox Proportional Hazards model (CoxPH) is a semi-parametric survival analysis model (Cox, 1972), consisting of a non-parametric baseline hazard function $h_0(t)$ and a parametric partial hazard function $f(\boldsymbol{\theta}, \boldsymbol{x}_i)$, typically estimated using the Breslow method (Breslow, 1975). The hazard function is $h(t \mid \boldsymbol{x}_i) = h_0(t)\exp(f(\boldsymbol{\theta}, \boldsymbol{x}_i))$, where f is linear. The MLE of $\widehat{\boldsymbol{\theta}}$ maximizes the Cox partial log-likelihood (Cox, 1972). CoxPH is implemented from the `scikit-survival` package (Pölsterl, 2020).

RSF: Random Survival Forests (RSF) (Ishwaran et al., 2008) extend decision trees for survival analysis by constructing multiple trees on bootstrapped samples. Each tree predicts an individual's survival function, with splits based on criteria like the log-rank test. RSF handles censored data as right-censored and accounts for it during tree construction, with final predictions as the average of all trees' survival functions. RSF is implemented from the `scikit-survival` package (Pölsterl, 2020).

DeepSurv: DeepSurv (Katzman et al., 2018) is a neural network implementation of the CoxPH model of the form $h(t \mid \boldsymbol{x}_i) = h_0(t)\exp(f(\boldsymbol{\omega}, \boldsymbol{x}_i))$, where $f(\boldsymbol{\omega}, \boldsymbol{x}_i)$ denotes a risk score given by a multilayer perceptron with Rectified Linear Unit (ReLU) activations and a linear output unit. The MLE for $\widehat{\boldsymbol{\omega}}$ is derived by numerically maximizing the partial Cox log-likelihood, with full-batch gradient steps because its risk set spans the dataset. Training uses Adam for at most 1000 epochs with early stopping at patience 10, and Breslow's estimator for survival curves. We follow the implementation of Katzman et al. (2018).

MTLR: Multi-Task Logistic Regression (MTLR) (Yu et al., 2011) is a discrete-time survival model that represents the event-time distribution through a sequence of dependent logistic regressors. We use a neural MTLR implementation with a dataset-specific number of time bins selected during hyperparameter tuning. The finite interval boundaries are defined by quantiles of the uncensored event times in the training set, such that each interval contains approximately the same number of observed events. An additional terminal interval represents event times beyond the final finite boundary. Right-censored observations contribute the probability of surviving beyond their censoring time. Our implementation of MTLR is adapted from the public implementation released with MENSA(Lillelund et al., 2026a).

Cox-Gamma Frailty: The Cox-Gamma frailty model extends proportional-hazards survival modelling with a subject-specific multiplicative frailty, such that $h(t \mid \boldsymbol{x}_i, u_i) = u_i h_0(t)\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$,

where $u_i \sim \mathrm{Gamma}(\alpha, \alpha)$ has unit mean. We use a piecewise-exponential baseline hazard and analytically integrate out the Gamma frailty to optimize the marginal likelihood. Population-level parameters are estimated using maximum a posteriori (MAP) estimation, while the subject-specific frailty posterior remains available in closed form through Gamma conjugacy. Thus, this is an empirical-Bayes frailty model rather than a full MCMC implementation. Our implementation uses the classical Gamma-frailty formulation (Vaupel et al., 1979; Hougaard, 1986) and follows the modelling construction illustrated in the PyMC frailty-model tutorial at `https://www.pymc.io/projects/examples/en/latest/survival_analysis/frailty_models.html`.

Clayton AFT: Following Foomani et al. (2023), Clayton AFT couples separate Weibull margins for the event time E and the censoring time C through a single-parameter Clayton copula (Clayton, 1978), fitting the margins and the dependence parameter jointly by maximizing the right-censored copula likelihood. Our margins are Weibull accelerated-failure-time models with a covariate-dependent log-scale, a reparameterization of their proportional-hazards margin for a linear predictor but without their neural risk function. We also apply the copula to the marginal distribution functions rather than, as they do, to the survival functions, so the fitted dependence is the survival ($180°$-rotated) Clayton – the same Kendall's τ family, with its asymmetry at early rather than late times. Since our generator couples the two times through their survival margins, the dependence family this baseline assumes is never the one that generated the data, in any scenario.

HACSurv: HACSurv is a neural copula-based survival model that jointly learns marginal survival distributions and their dependence structure (Liu et al., 2025). Its marginal survival functions are represented by monotone neural networks, and dependence is represented using an Archimedean-copula generator parameterized as a stochastic mixture of exponentials. Although the original method accommodates hierarchical dependence among competing risks and censoring, our setting contains one event time and one censoring time; we therefore use its bivariate special case. In this case HACSurv, like DCSurvival (Zhang et al., 2024), learns an Archimedean generator jointly with the margins; we therefore use HACSurv as the representative learned-copula baseline and do not evaluate DCSurvival separately. HACSurv is fitted by maximizing the bivariate observed-data likelihood and provides a flexible dependent-censoring comparator without imposing parametric event or censoring margins. We use the implementation at `https://github.com/Raymvp/HACSurv`.

G.5 Performance metrics

Oracle prediction metrics. In both studies the data-generating process supplies the true event time e_i for every subject, including those whose observation is censored. We therefore evaluate on the complete sample, replacing (t_i, δ_i) by $(e_i, \delta_i \equiv 1)$ throughout, and adopt the oracle formulations of the C-index, IBS and MAE defined by Lillelund et al. (2026c). This removes the inverse-probability weights $\widehat{G}$ (Graf et al., 1999) that their censored counterparts require and makes every pair comparable, so the oracle metrics carry no dependence on a censoring model.

All three are computed on one grid $\mathcal{G}$ of 100 points spanning $[0, t_{\max}]$, shared by every model and by both studies, with $t_{\max}$ the 95th percentile of the true event times in the training split. Anchoring the horizon on observed times instead would truncate it by exactly the censoring these metrics are defined to see past: the observed event times belong to the uncensored subjects, who are the fast-failing minority once the censoring rate is high, so the horizon would be shortest on the datasets that need it longest. The quantile is taken on the training split alone, so no held-out outcome enters the grid.

Where a point prediction is required, we use the predicted median event time

$$\widehat{e}_i = \begin{cases} \min\{t \in \mathcal{G} : \widehat{S}_E(t \mid \boldsymbol{x}_i) \le 0.5\}, & \text{if the curve reaches } 0.5 \text{ on } \mathcal{G}, \\ t_{\max}, & \text{otherwise.} \end{cases} \tag{39}$$

A curve that never reaches 0.5 has no median identified within the horizon, so we clip at $t_{\max}$ rather than extrapolate beyond the grid. This keeps every point prediction inside the region on which the curves were evaluated and $\mathrm{MAE}_{\mathrm{oracle}}$ bounded; its cost is that such subjects tie at $t_{\max}$.

Oracle IBS. The Brier score at a fixed time t^* is the unweighted squared error against the true survival indicator,

$$\mathrm{BS}_{\mathrm{oracle}}(t^*) = \frac{1}{N} \sum_{i=1}^{N} \Big(\mathbb{1}[e_i > t^*] - \widehat{S}_E(t^* \mid \boldsymbol{x}_i) \Big)^2 . \tag{40}$$

and the integrated score averages it over the grid,

$$\mathrm{IBS}_{\mathrm{oracle}} = \frac{1}{t_{\max}} \int_0^{t_{\max}} \mathrm{BS}_{\mathrm{oracle}}(t)\, \mathrm{d}t. \tag{41}$$

We evaluate the integral by trapezoidal quadrature.

Oracle C-index. Harrell's concordance between the true event times and the predicted medians,

$$C_{\mathrm{oracle}} = \frac{\sum_{i,j} \mathbb{1}[e_i < e_j] \left(\mathbb{1}[\widehat{e}_i < \widehat{e}_j] + \frac{1}{2} \mathbb{1}[\widehat{e}_i = \widehat{e}_j] \right)}{\sum_{i,j} \mathbb{1}[e_i < e_j]}. \tag{42}$$

with tied predictions contributing $1/2$, which is common here because (39) quantizes to the grid and clips. Every pair with distinct event times, (i, j) is comparable, so the δ_i factor of the censored estimator drops out and no weighting is required.

Oracle MAE. The mean absolute error between the true event times and the predicted medians,

$$\mathrm{MAE}_{\mathrm{oracle}} = \frac{1}{N} \sum_{i=1}^{N} \left| e_i - \widehat{e}_i \right|. \tag{43}$$

Lower is better for oracle IBS and MAE, higher for the oracle C-index. Oracle IBS is the metric used for hyperparameter selection.

Shared-latent recovery. Because the shared latent z is generated by the DGP, it can be compared directly against what the model infers. For each held-out subject we take the mean of the encoder's approximate posterior, $\widehat{z}_i$, and report the Spearman rank correlation between $\{\widehat{z}_i\}$ and the true $\{z_i\}$. A rank correlation is appropriate because the latent's location and scale are not identified, so only the ordering it induces over subjects is meaningful; the reported value is the raw correlation, with no sign correction applied, and it is positive in every run we report. This metric is defined only for models that infer a subject-level latent, and only where the DGP carries a shared-frailty signal: at $\tau = 0$ the latent is unidentified and recovery is undefined rather than zero.

Conditional Kendall's τ error. The DGP induces a known rank association between the event and censoring times. We compare it against the association implied by a fitted model, evaluated *conditional on covariates* at a reference vector $\boldsymbol{x}$ taken as the training-set mean, since the dependence a model represents may vary with $\boldsymbol{x}$ and pooling subjects with different covariates would confound it with covariate-driven shifts in the marginals. Writing $\tau^\star$ for the target and $\widehat{\tau}(\boldsymbol{x})$ for the model's estimate, we report $|\widehat{\tau}(\boldsymbol{x}) - \tau^\star|$. For DVFM, $\widehat{\tau}(\boldsymbol{x})$ is the empirical Kendall's τ of 2,000 pairs (E, C) sampled from the fitted decoder at $\boldsymbol{x}$; for HACSurv, it is computed analytically from the fitted Archimedean generator φ, taken in the Laplace-transform parameterization (φ completely monotone, $\varphi(0) = 1$, $\varphi(s) \to 0$ as $s \to \infty$), as

$$\tau = 1 - 4 \int_0^{\infty} s\, \varphi'(s)^2\, \mathrm{d}s.$$

Thus, the DVFM estimate is subject to Monte Carlo error, whereas the HACSurv estimate is exact up to numerical quadrature.

Joint-survival integrated squared error (ISE). Standard survival evaluation considers the marginal event-time survival function $\Pr(E > e \mid \boldsymbol{X})$. Under dependent censoring, this does not assess whether a model captures the dependence between the event and censoring processes. We therefore evaluate the conditional joint survival function

$$S(e, c \mid \boldsymbol{x}) = \Pr(E > e, C > c \mid \boldsymbol{X} = \boldsymbol{x}). \tag{44}$$

Its boundaries recover the event- and censoring-time marginals, while its interior characterizes their dependence. Unlike a scalar association measure such as Kendall's τ, joint-survival ISE evaluates the entire predicted joint survival surface.

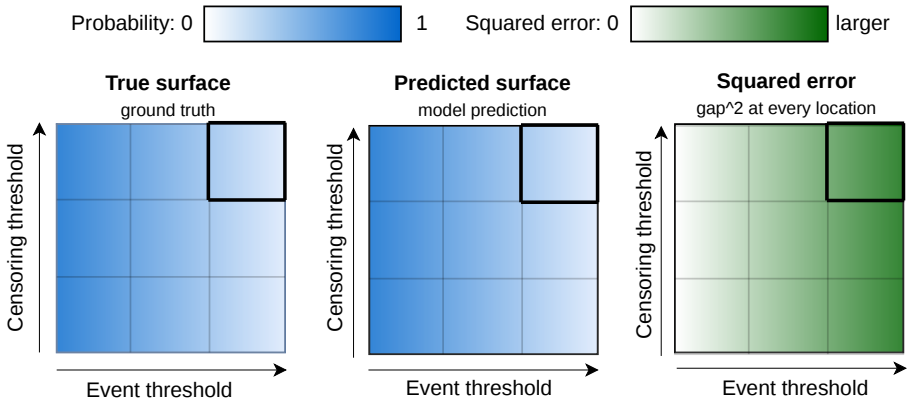


Figure 10: Illustration of joint-survival ISE. The left and center panels show the true and predicted joint survival probabilities at each threshold pair (e, c). The right panel shows their pointwise squared difference, which is averaged over the grid and held-out subjects. Its color scale is adjusted separately for visibility.

Let $\widehat{S}(e, c \mid \boldsymbol{x})$ denote the estimated joint survival function. For N held-out subjects, we define

$$\mathrm{ISE} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{e_{\max} c_{\max}} \int_0^{e_{\max}} \int_0^{c_{\max}} \left(\widehat{S}(e, c \mid \boldsymbol{x}_i) - S(e, c \mid \boldsymbol{x}_i)\right)^2 \mathrm{d}c\,\mathrm{d}e, \tag{45}$$

where $e_{\max}$ and $c_{\max}$ are the 95th percentiles of the true event and censoring times in the training split. Normalization by $e_{\max} c_{\max}$ yields a mean squared error over the evaluation rectangle, making the metric invariant to multiplicative rescaling of either time axis. Thus, ISE $\in [0, 1]$, and $\sqrt{\mathrm{ISE}}$ is the root-mean-square error in the predicted joint probability. Lower values are better. Figure 10 illustrates the metric.

In practice, we approximate the integrals using trapezoidal quadrature on a 15×15 grid for $N = 128$ held-out subjects. The reference surface is estimated from 2,000 draws from the known DGP, whereas $\widehat{S}$ is evaluated in closed form or estimated from 500 posterior samples, depending on the model.

Held-out log-likelihood. Without ground truth, the real-data experiment scores the observed outcome directly, as the mean over held-out patients of $\log p_\theta(t, \delta \mid \boldsymbol{X})$, which requires neither event times nor independent censoring. For d_z=0 it is exact. For d_z=1 we estimate it by importance sampling with the encoder (6) as proposal and K = 1,000 draws (Burda et al., 2016), a lower bound that reduces to the ELBO (8) at K=1. Follow-up is administratively censored at h = five years: a patient free of both outcomes at h contributes $\log \Pr(E > h, C > h \mid \boldsymbol{X})$ instead of a density, which affects at most four test patients per split. The death and censoring margins integrate each margin over the aggregate posterior (11) and score it as if censoring were independent, so they are descriptive and do not sum to the joint.

G.6 Performance reporting protocol

This appendix states how per-run metric values become the ranks and intervals reported in Figure 4 and Table 1, following Qi et al. (2026).

Runs. The benchmark comprises 8 models $\times$ 12 datasets $\times$ 5 scenarios $\times$ 10 seeds = 4,800 runs, plus the 600 runs of DVFM's matched d_z=0 ablation that Figure 4 ranks alongside them, the scenarios being independence at $\tau = 0$ and the four copulas at $\tau = 0.5$ (Section 6). Seed $r \in \{0, \ldots, 9\}$ is passed independently to dataset generation, copula sampling, splitting and model initialization, and every model within a cell is fitted to the identical $70/10/20$ split, stratified jointly by event status and observed-time quantile band, so comparisons are paired. Hyperparameters are fixed beforehand on a development split that this evaluation never reuses (Appendix G.8).

Ranking. The procedure below is applied separately to each of oracle CI, IBS and MAE; we drop the metric index for readability. Let $\mathrm{rank}_{m,D,c}$ denote the rank of model m on dataset D in evaluation cell c, taken over the nine models sharing that cell, where a cell is one copula-seed combination, so $|\mathcal{C}| = 40$ at $\tau = 0.5$ and $|\mathcal{C}| = 10$ at $\tau = 0$. Rank 1 is assigned to the best value, with ties receiving the minimum tied rank. A non-finite metric value or a run flagged as a numerical failure is assigned one rank below the worst completed model in the same cell, penalizing the failure without displacing the models that completed. No run triggered this rule: all $5{,}400$ produced finite oracle CI, IBS and MAE, a failure rate of 0%

Aggregation. Ranks are averaged over the cells of a dataset,

$$\overline{\mathrm{rank}}_{m,D} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathrm{rank}_{m,D,c}, \tag{46}$$

which weights the four copulas equally because the copula-seed grid is complete. Each model is then summarized by the median over the 12 datasets, $\widetilde{r}_m = \mathrm{median}_D \, \overline{\mathrm{rank}}_{m,D}$.

Uncertainty. Seeds and scenarios are averaged out before any interval is computed, so the unit of analysis is the dataset and the population generalized over is that of datasets, not the seed noise within one. We therefore bootstrap over datasets: each replicate $b = 1, \ldots, B$ draws 12 datasets with replacement and recomputes $\widetilde{r}^{(b)}_{m,q}$, and we report the 95% percentile interval $[Q_{0.025}\{\widetilde{r}^{(b)}_{m,q}\}^B_{b=1},\ Q_{0.975}\{\widetilde{r}^{(b)}_{m,q}\}^B_{b=1}]$ with $B = 10{,}000$ under a fixed seed. Here B controls Monte Carlo error only and is not a sample size; with 12 datasets the intervals stay wide and coarse, a property of the benchmark's size rather than of the estimator, and we read overlapping intervals as indistinguishable performance rather than as a ranking.

Applicability. Oracle CI, IBS and MAE are defined for all nine models, whereas the joint-survival ISE requires a predicted bivariate surface and is reported only for DVFM and HACSurv. Oracle CI and MAE are computed from the predicted median event time (39), which is clipped at $t_{\max}$ for any curve that does not reach 0.5 within the horizon; such subjects tie at $t_{\max}$, and the oracle C-index credits a tied pair $1/2$.

G.7 Implementation details

Architecture. The encoder is a multilayer perceptron over $(\boldsymbol{x}, t, \delta)$ – covariates together with the observed follow-up time and the event indicator – with ReLU activations and batch normalization, and two linear heads emitting the posterior mean and log-variance of z. The log-variance is clamped to $[-12, 8]$, which bounds the posterior standard deviation $\exp(\frac{1}{2}\log\sigma^2)$ during both training and Monte Carlo prediction so that an extreme held-out subject cannot overflow. The decoder is a multilayer perceptron over $(\boldsymbol{x}, z)$ with the same activation pattern and a single linear head emitting four values, mapped to the Weibull shape and scale parameters of the event and censoring margins: shapes through a softplus and scales through an exponential link whose argument is clamped to $[-12, 12]$, both offset by a 10^{-6} floor. The scale link is a configuration option (softplus or exponential), fixed to exponential in all reported runs. The shape is subject-specific, so the model does not impose proportional hazards; a "global" shape mode that does impose it is implemented and available in the configuration schema, but is not exercised in the reported experiments.

Objective. Training DVFM maximizes the β-regularized censored ELBO of Section 3. The reconstruction term is the per-subject mean of $\delta(\log f_E + \log S_C) + (1-\delta)(\log S_E + \log f_C)$, so a censored subject contributes an observed censoring event rather than being discarded; the KL term is summed over latent dimensions and averaged over subjects. The weight β increases linearly from $1/\text{warm-up}$ to $\beta_{\max} = 1$ over 50 warm-up epochs and is held there, and free bits are disabled (0 nats).

Latent dimension. All reported runs use $d_z = 1$, chosen so that the inferred frailty is a single interpretable quantity comparable against the generating latent; larger d_z would forfeit that diagnostic, which is the study's main evidence that the shared pathway does what it claims. The matched $d_z = 0$ control removes the encoder and the latent path while sharing every architecture and optimization setting, so the two differ only in the shared-latent pathway. It is fitted in all three studies: as the paired control in the synthetic study and the real-data experiment, and as a ranked entry in the semi-synthetic pool.

Latent-loading shrinkage. An ℓ_1 penalty with $\alpha = 0.1$, distinct from the DGP loading λ, is applied to the mean absolute value of the decoder weights carrying z into the margin heads. It shrinks how strongly the latent may act on the decoded margins rather than the magnitude of z itself, which is the quantity entering the covariance of Proposition 1. It is on by default in every reported run. The value was selected in a prespecified sensitivity analysis on separate synthetic datasets at 25% censoring, comparing an unregularized reference against loading-ℓ_1 variants at $\{0.001, 0.01, 0.1\}$ across $\tau \in \{0, 0.25, 0.5, 0.75\}$, scored by oracle IBS on the validation partition under prior sampling; a gating alternative was evaluated in the same analysis and not retained. The configuration is in the supplementary material. α is then held constant across every dataset and scenario rather than tuned.

Optimization. For the synthetic experiments, we use Adam, learning rate 10^{-3}, batch size 64, 200 epochs, weight decay 10^{-4}, and gradient-norm clipping at 1.0, in single precision. For the semi-synthetic experiments, learning rate, weight decay, dropout and architecture are tuned per dataset. The remaining settings are fixed throughout. For the real-data experiment, both models use a $[64, 32]$ encoder and a $[32, 64]$ decoder, learning rate 10^{-3}, weight decay 10^{-5} and no dropout, set by hand rather than tuned, since without an oracle there is no tuning criterion independent of the evaluation.

Checkpointing and numerical guards. The reported model is the checkpoint with the best validation ELBO after warm-up; no test data influences selection. A non-finite objective or gradient norm aborts the run, and epochs whose absolute ELBO, reconstruction negative log-likelihood or KL exceeds 100 are rejected as numerical failures. Finite predictions alone are not treated as evidence that a fit succeeded.

Prediction. The latent is unobserved for a new subject, so survival curves are obtained by Monte Carlo integration over the aggregate training posterior with 100 samples. We use the aggregate posterior rather than the prior because it estimates the model's own marginal predictive directly, whereas prior sampling is valid only insofar as the aggregate posterior matches $\mathcal{N}(0, I)$. Prior sampling is supported by the implementation but is not used in any reported row.

G.8 Hyperparameter tuning

We tuned hyperparameters separately for each dataset on a development-only dataset. The dataset was generated with seed 99, a Gaussian copula, Kendall's $\tau = 0.50$, and the dataset's observed censoring rate, then split once into 70% training and 30% validation, stratified jointly by event status and observed-time bin. No test partition was used during tuning, and neither the dataset nor seed 99 was reused in the final 10-seed evaluation.

CoxPH was used with implementation defaults. For every other learner, we independently sampled $R = 10$ configurations from the model-specific discrete candidate set $\mathcal{C}_m$ in Table 5. Each candidate was fitted on the development training partition and scored by its oracle integrated Brier score (IBS) on the validation partition. The selected configuration was

$$c_m^* = \arg \min_{c \in \{c_1, \ldots, c_R\} \subset \mathcal{C}_m} \mathrm{IBS}_{\mathrm{oracle}}^{\mathrm{val}}(c; m).$$

Oracle IBS is available because the semi-synthetic DGP supplies the true event-time survival distribution. Failed configurations were treated as infeasible and excluded from selection. The selected settings were then fixed for all final scenarios and repeats. DVFM was tuned with a one-dimensional latent variable ($d_z = 1$), and the resulting architecture and optimization settings were shared with its $d_z = 0$ ablation.

All candidates used the implementation's remaining fixed settings. In particular, DeepSurv, MTLR, and Clayton AFT used their configured fixed epoch budgets, Clayton AFT's being 100 full-batch Adam epochs; HACSurv-2D and Cox-Gamma frailty used validation-based early stopping; and DVFM trained for its fixed epoch budget while retaining the best eligible post-warm-up validation-ELBO checkpoint.

Table 5: Discrete hyperparameter search spaces for the semi-synthetic experiments. Each listed choice is sampled independently, except that a DVFM architecture is sampled as an encoder-decoder pair and HACSurv's copula learning rate is the sampled multiplier times the sampled learning rate.

Model	Hyperparameter	Candidate values
DeepSurv	learning rate	$\{10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 10^{-1}\}$
	weight decay	$\{0, 10^{-5}, 10^{-4}, 10^{-3}\}$
	hidden layers	$\{[16], [32, 16], [64], [64, 32], [64, 64, 16]\}$
	dropout	$\{0, 0.1, 0.25\}$
MTLR	learning rate	$\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$
	weight decay	$\{0, 10^{-5}, 10^{-4}, 10^{-3}\}$
	hidden layers	$\{[16], [32, 16], [64], [64, 32]\}$
	dropout	$\{0, 0.1, 0.25\}$
	time bins	$\{50, 100, 200\}$
RSF	number of trees	$\{100, 300, 500\}$
	maximum depth	$\{3, 5, 8\}$
	minimum split size	$\{2, 10, 25\}$
	minimum leaf size	$\{1, 5, 10\}$
	maximum features	$\{\texttt{sqrt}, \texttt{log2}, 0.5\}$
Clayton AFT	learning rate	$\{10^{-4}, 5\times10^{-4}, 10^{-3}, 5\times10^{-3}, 10^{-2}\}$
HACSurv	learning rate	$\{3\times10^{-5}, 10^{-4}, 3\times10^{-4}\}$
	copula learning-rate multiplier	$\{0.3, 1, 3\}$
	marginal-network width	$\{16, 32, 64\}$
	survival-network width	$\{16, 32, 64\}$
	scale regularization	$\{0.1, 1, 10\}$
	weight decay	$\{0, 10^{-5}, 10^{-4}, 10^{-3}\}$
Cox-Gamma frailty	learning rate	$\{0.003, 0.01, 0.03, 0.1\}$
	baseline-hazard intervals	$\{5, 10, 20\}$
	baseline smoothness	$\{0.1, 1, 10\}$
	regression-coefficient prior SD	$\{1, 2.5, 5\}$
DVFM	learning rate	$\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}\}$
	weight decay	$\{0, 10^{-5}, 10^{-4}, 10^{-3}\}$
	encoder-decoder pairs	$([32, 16], [16, 32])$ $([64, 32], [32, 64])$ $([64, 64, 16], [16, 64, 64])$
	dropout	$\{0, 0.1, 0.25\}$

H Additional results

H.1 Absolute metric values behind the ranks

Tables 6 and 7 report the oracle metrics for the nine models, along with relative changes against DVFM. Seeds are averaged within each dataset–copula cell, then copulas within each dataset. The reported value is the median across datasets; Δ is the median across datasets of the within-dataset relative change, $(\text{model} - \text{DVFM})/|\text{DVFM}|$. Because these medians are calculated separately, Δ need not equal the relative difference between the displayed values. Oracle MAE retains each dataset's time scale, so its absolute values should be interpreted with that limitation.

The ranks conceal substantial differences in separation between metrics. At $\tau = 0.5$, the reported oracle CI values lie between 0.643 and 0.677, and the median within-dataset changes against DVFM have magnitudes of at most 2.2%. By contrast, the four independent-censoring models without a frailty term have median within-dataset increases of 21.5%-34.7% in oracle IBS or MAE, while Cox-Gamma frailty stays within 1.5% of DVFM on both. Dependence also widens the spread in oracle IBS: the reported changes range from -6.7% to $+5.6\%$ at $\tau = 0$, compared with -1.3% to $+34.7\%$ at $\tau = 0.5$. Lower IBS and MAE are better; higher CI is better.

Importantly, the ablation's median understates its spread. Tables 6 and 7 place d_z=0 close to DVFM at both dependence levels, but the per-dataset differences are large and of both signs. Relative to

Table 6: Oracle metrics at Kendall's $\tau = 0$. Values and changes are defined in the text.

	Oracle IBS		Oracle CI		Oracle MAE	
Model	Value	Δ	Value	Δ	Value	Δ
CG Frailty	0.106	-6.7%	0.668	+1.6%	406	-3.3%
DVFM (d_z=0)	0.116	-1.0%	0.668	-0.1%	415	-0.3%
DVFM	0.120	–	0.657	–	422	–
HACSurv	0.122	-1.6%	0.665	+0.8%	411	+0.6%
Clayton AFT	0.125	+3.9%	0.652	-0.3%	427	+0.9%
CoxPH	0.126	-2.0%	0.652	+1.0%	461	-0.7%
MTLR	0.126	+4.6%	0.655	-1.2%	438	+4.1%
DeepSurv	0.131	+3.3%	0.653	+0.2%	537	+1.8%
RSF	0.134	+5.6%	0.645	-1.6%	481	+2.6%

Table 7: Oracle metrics at Kendall's $\tau = 0.5$. Values and changes are defined in the text.

	Oracle IBS		Oracle CI		Oracle MAE	
Model	Value	Δ	Value	Δ	Value	Δ
DVFM	0.111	–	0.665	–	415	–
HACSurv	0.116	-1.3%	0.668	+0.3%	438	-0.8%
CG Frailty	0.121	-0.5%	0.671	+0.8%	406	+1.4%
DVFM (d_z=0)	0.124	+1.3%	0.677	-0.2%	407	+0.4%
Clayton AFT	0.125	+2.7%	0.660	-0.5%	425	+1.1%
MTLR	0.146	+28.5%	0.646	-1.5%	553	+21.5%
RSF	0.152	+32.4%	0.643	-2.2%	560	+26.4%
CoxPH	0.161	+28.9%	0.654	+0.1%	586	+28.5%
DeepSurv	0.165	+34.7%	0.649	-1.1%	605	+32.5%

the ablation, adding the latent changes oracle IBS within a dataset by -25.0% to $+8.2\%$ at $\tau = 0.5$: it lowers IBS by 10.6%-25.0% on Churn, Employee, MIMIC-IV and NACD, raises it by 7.9% and 8.2% on SEER (liver) and SEER (brain), and moves it by less than 3% on the remaining six datasets. DVFM has the lower oracle IBS on 7 of the 12 datasets at $\tau = 0.5$ and on 5 of 12 at $\tau = 0$, where the median change reverses sign to $+1.0\%$. The latent's value is therefore dataset-dependent rather than uniformly small.

H.2 Latent recovery across copula families

Figure 11 repeats the subject-level diagnostic on the SUPPORT dataset at $\tau = 0.5$ under each of the four dependence families, pooling held-out subjects over the ten repeats; hexagonal binning replaces a scatter because at this sample size overplotting hides the density. Each repeat keeps its own affine calibration, learned on that repeat's validation split. The recovered latent tracks the generating variable under every family, so recovery is not an artifact of the Gaussian construction used in Figure 1. The estimand differs by family – the common normal factor for the Gaussian copula, the standardized log mixing variable for the Archimedean ones – so these panels compare recovery quality, not a shared quantity, and we do not read an ordering between families from them: the spread across families is far smaller than the spread across datasets.

H.3 Dependence calibration

Figure 12 shows that DVFM detects dependence but does not measure it. The learned conditional τ increases from the $\tau = 0$ to the $\tau = 0.5$ condition on almost every dataset, but by at most about 0.15, and each dataset carries an offset shared by both targets that grows with the number of uncensored training events, reaching roughly 0.36 at a true τ of 0. A single estimate near its target is therefore not evidence of calibration: SUPPORT reaches 0.47 at $\tau = 0.5$ while the same model reports 0.34 at $\tau = 0$. The limitation is not specific to any dependence geometry, since the error varies far more across datasets than across copula families (Figure 12a). DVFM thus recovers the presence and direction of dependence, more strongly on larger datasets, but the estimate's level is not comparable across datasets and a single value carries no magnitude information.

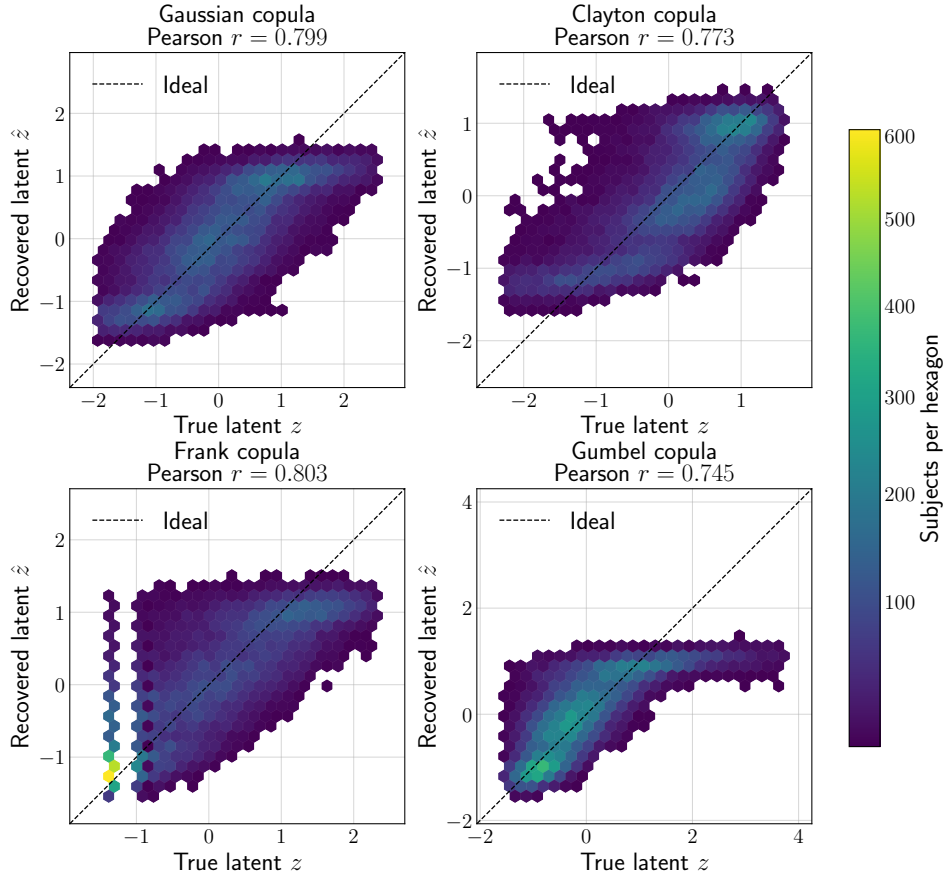
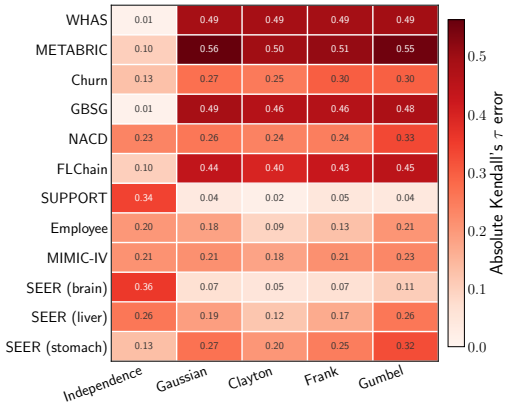
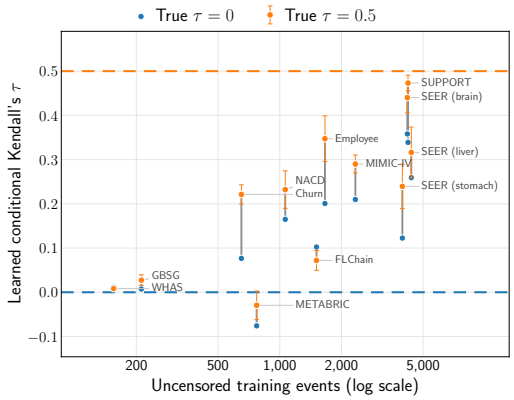


Figure 11: Subject-level latent recovery on the SUPPORT dataset at $\tau = 0.5$ under each dependence family, pooling held-out subjects over the ten repeats. Hexagon shading gives the number of subjects per bin on a shared scale, the dashed line marks perfect recovery, and each panel title reports the Pearson correlation for that family. The recovered latent is affinely calibrated within each repeat on that repeat's validation split.



(a) Absolute τ error by dataset and copula.



(b) Learned τ vs. uncensored training events.

Figure 12: **Dependence is detected, but its magnitude is not recovered.** (a) Absolute error $|\hat{\tau} - \tau|$ per dataset $\times$ copula; it varies far more across datasets than across copula families. (b) Learned conditional Kendall's τ against uncensored training events, with the two targets as dashed lines and a stub joining each dataset's pair. The estimate moves in the right direction on almost every dataset, but by at most about 0.15 for a true change of 0.5, and both targets share a dataset-specific offset that grows with the event count.