TESTING DEPENDENT CENSORING VIA MUTUAL-INFORMATION UNCERTAINTY AND DEEP LATENT IMPUTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning a survival prediction model can be viewed as regression with the added complication of censoring. Each subject x_i has a true event time E_i and a censoring time C_i, yet we only observe $T_i = \min(E_i, C_i)$ and $\delta_i = \mathbf{1}(E_i \leq C_i)$. Many standard methods assume conditional independence $E \perp C \mid \mathbf{X}$, which may not hold in practice. Assessing this dependence is challenging, as for each subject we observe either E_i or C_i, but never both. To address this, for each $t > 0$ we define indicator variables $\tilde{E}_i(t) = \mathbf{1}(E_i > t)$ and $\tilde{C}_i(t) = \mathbf{1}(C_i > t)$, taking values in $\{0, 1, ?\}$, where "?" denotes censoring-induced missingness. Using $\{\tilde{E}_i(t), \tilde{C}_i(t)\}_t$ across multiple time points, we develop a model-assisted diagnostic for testing $E \perp C \mid \mathbf{X}$. Our statistic is the Conditional Mutual Information (CMI) uncertainty range, $\Delta I_t = I^t_{\max} - I^t_{\min}$, defined as the width of the CMI values attainable over all feasible completions of the missing "?" entries. Here, $I^t_{\max}$ and $I^t_{\min}$ are the corresponding maximum and minimum feasible CMI values. Under independence, ΔI_t follows a characteristic null distribution, whereas dependent censoring induces structured deviations. To compute these bounds efficiently, we formulate the problem as a decomposable integer program and solve it exactly via a polynomial-time algorithm. To calibrate the statistic, we combine this uncertainty measure with a deep variational frailty model that imputes latent event and censoring marginals, followed by stratified permutation to generate a model-assisted null distribution under conditional independence. Synthetic experiments show calibrated behavior under independent censoring and increasing power for several dependent-censoring mechanisms, while real-data experiments reveal heterogeneous evidence of possible dependent censoring.

1 INTRODUCTION

Survival analysis models time-to-event data in applications ranging from clinical prognosis to engineering reliability (Clark et al., 2003). A central difficulty is right censoring: for each subject, the true event time may be unobserved because follow-up ends before the event occurs. Standard survival methods, including the Cox proportional hazards model (Cox, 1972) and modern neural approaches such as MTLR (Yu et al., 2011) and DeepSurv (Katzman et al., 2018), typically rely on the independent censoring assumption. This assumption states that, conditional on covariates $\mathbf{X}$, the latent event time E is independent of the latent censoring time C.

In practice, this assumption may fail. For example, in a clinical trial, patients with worsening health may leave the study early to seek alternative treatment, making early censoring informative about event risk. Such dependent censoring can bias survival estimates and lead to invalid conclusions (Lillelund et al., 2026; Gharari et al., 2023). Detecting this violation is difficult because E and C are never jointly observed: for each subject, we only observe $T = \min(E, C)$ and $\delta = \mathbf{1}\{E \leq C\}$.

1.1 THE TSIATIS NONIDENTIFIABILITY CHALLENGE

The difficulty of testing independent censoring is formalized by the Tsiatis nonidentifiability result (Tsiatis, 1975). The observed data can be consistent with both a truly dependent joint distribution of

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.

(E, C) and an independent proxy distribution that yields the same distribution of (T, δ). Consequently, likelihood-based methods cannot generally distinguish dependent censoring from its independent proxy using observed data alone. This motivates the need for additional structure that can expose dependence beyond what is identifiable from the observed likelihood.

1.2 Existing Approaches

Existing approaches address dependent censoring from several directions. Cox-model-based tests using additional follow-up information assess independent censoring under proportional hazards assumptions (Lee and Wolfe, 1998). Joint modeling, frailty, and copula-based approaches explicitly model the association between event and censoring processes, but require assumptions on the dependence structure or joint distribution, which may lead to misspecification (Rizopoulos, 2012). Rank-based tests using additional follow-up information (Lee and Wolfe, 1995) and tests for longitudinal responses (Sun and Lee, 2011) address related censoring questions. Weighted log-rank methods instead compare survival functions in the presence of dependent censoring (Flandre, 2022). Sensitivity analysis provides another perspective by quantifying how conclusions change under possible violations of independent censoring, rather than directly testing for dependence (Scharfstein et al., 1999).

Despite these advances, existing methods often rely on restrictive assumptions, have limited ability to detect complex dependence, or do not directly address the Tsiatis nonidentifiability barrier. This motivates a testing framework tailored to censored data, based on information-theoretic uncertainty over feasible latent completions and structural generative modeling.

1.3 Our Contributions

This paper develops a model-assisted diagnostic for dependent censoring based on mutual-information uncertainty. The main contributions are:

1. **Discrete-time censoring representation.** We represent latent event and censoring times through landmark indicators $E_t = \mathbf{1}\{E > t\}$ and $C_t = \mathbf{1}\{C > t\}$. Continuous-time conditional independence implies conditional independence of these indicators at every landmark time, which provides a testable family of necessary conditions.

2. **Mutual-information uncertainty statistic.** Because censoring makes the binary pair (E_t, C_t) partially observed, the empirical conditional mutual information is not uniquely determined. We define the MI uncertainty range ΔI_t as the width between the minimum and maximum feasible conditional mutual information over all completions consistent with the observed data.

3. **Exact stratum-wise computation.** We show that the feasible-completion problem decomposes across covariate strata. This yields an exact finite optimization over two integer variables per stratum, avoiding enumeration over all subject-level completions.

4. **Model-assisted permutation calibration.** We use a Deep Variational Frailty Model (DVFM) to estimate latent event and censoring marginals, then apply stratified permutation of imputed censoring times to generate a reference distribution under conditional independence. This produces a calibrated diagnostic under explicit modeling assumptions.

Together, these components provide a practical way to detect identifiable departures from independent censoring while making clear which part of the procedure is data-driven and which part relies on model-assisted latent imputation.

2 Background: Survival Analysis and Independent Censoring

Survival analysis models the time until an event of interest occurs. In the presence of right censoring, each subject has a latent event time E and a latent censoring time C, but only the observed follow-up time and event indicator are available:

$$T = \min(E, C), \qquad \delta = \mathbf{1}\{E \leq C\}, \tag{1}$$

where $\delta = 1$ indicates an observed event and $\delta = 0$ indicates right censoring. The observed data for each subject are therefore $(T, \delta, \mathbf{X})$, where $\mathbf{X}$ denotes the covariates.

Given covariates $\mathbf{X}$, the latent event process is characterized by the survival function $S_E(t \mid \mathbf{X}) = P(E > t \mid \mathbf{X})$, and the marginal hazard function

$$h_E(t \mid \mathbf{X}) = \lim_{\Delta t \to 0} \frac{P(t \leq E < t + \Delta t \mid E \geq t, \mathbf{X})}{\Delta t} = -\frac{d}{dt} \log S_E(t \mid \mathbf{X}).$$

A central assumption in standard survival methods, including the Kaplan–Meier estimator and the Cox proportional hazards model (Cox, 1972), is independent censoring:

$$E \perp C \mid \mathbf{X}. \tag{2}$$

Under this assumption, censoring carries no additional information about the event time beyond the covariates.

When $E \not\perp C \mid \mathbf{X}$, censoring is dependent. In this case, the observed hazards may be distorted versions of the true latent hazards, leading to biased estimation and invalid inference. Detecting such dependent censoring is therefore the central objective of this work.

3 From Continuous Processes to Discrete-Time Representations

To test the independence hypothesis $E \perp C \mid \mathbf{X}$, we map the continuous-time variables to a discrete-time representation. This discretization is to transform an otherwise intractable problem into a tractable one. In particular, it enables the computation of the range of CMI between event and censoring processes over all feasible imputations of the unobserved values induced by censoring. By representing the process through binary indicators at multiple time points, we obtain a discrete structure that supports efficient evaluation of these CMI bounds.

3.1 Discrete-Time Representation of Survival Data

To discretize the continuous process, we select K ordered landmarks, either fixed independently of the outcomes or obtained from a prespecified quantile rule applied separately to the original and each reconstructed dataset. Data-adaptive landmarks are placed between adjacent distinct observed times to avoid equality with an observed time. Write the grid as $\mathcal{T} = \{t_1, \ldots, t_K\}$. At each time point $t_k \in \mathcal{T}$, we define two binary random variables for each subject:

- $E_{t_k} := \mathbf{1}\{E > t_k\}$, the **survival status indicator**. $E_{t_k} = 1$ if the subject has survived past time t_k.
- $C_{t_k} := \mathbf{1}\{C > t_k\}$, the **censoring status indicator**. $C_{t_k} = 1$ if the subject has not been censored by time t_k.

A critical aspect of this representation is the induced missingness. After a subject is censored at time C, their true event status E_t for all $t > C$ becomes unobserved. Similarly, once an event occurs at time E, the subsequent censoring status C_t for $t > E$ is not observed, as the subject is no longer in the risk set. Table 5 and Fig. 1 in Appendix D illustrate this transformation, showing the discrete-time trajectories for four subjects based on their observed event xor censoring times. As Subject A experiences the event at $t = 1.5$: their $[E_t, C_t]$ transition from $[1, 1]$ before $t = 1.5$, to become $[0, ?]$ afterwards. For Subject C, censored at $t = 3$, the survival status E_t is known to be 1 up to that point but becomes unobserved ? thereafter and C_t goes from 1 to 0.

3.2 From Continuous-Time to Landmark Independence

A crucial question is whether the independent censoring assumption from the continuous domain translates to the discrete representation. We demonstrate that it does, which provides the theoretical justification for our testing framework.

Proposition 1 *Let E and C be continuous event and censoring times, and let $\mathbf{X}$ be a vector of covariates. The continuous-time independent censoring assumption, $H_0^C : E \perp C \mid \mathbf{X}$, implies the discrete-time independent censoring assumption,*

$$\forall t > 0, \quad H_0^D(t) : E_t \perp C_t \mid \mathbf{X}, \tag{3}$$

where $E_t = \mathbf{1}\{E > t\}$ and $C_t = \mathbf{1}\{C > t\}$.

This proposition is the cornerstone of our testing methodology. Its contrapositive guarantees that **if there exist a t such that $H_0^D(t)$ fails, then H_0^C fails**. This insight transforms the testing problem: we need only find a single, statistically significant instance of discrete dependence to reject the global null hypothesis. Accordingly, our global test is designed to find the maximum deviation from independence across a set of discrete time points, while correcting for the multiple comparisons involved in such a search.

4 Introduction of an Information-Theoretic Statistic test

4.1 Quantifying Dependence with Conditional Mutual Information.

The core of our test is the **CMI**, a fundamental measure of the strength of dependence between two random variables given a third. For a chosen time point t, we measure the dependence between the survival status E_t and the censoring status C_t conditioned on the covariates $\mathbf{X}$. The CMI is defined, for any fixed t, as:

$$I(E_t; C_t \mid \mathbf{X}) = \mathbb{E}_{p(E_t, C_t, \mathbf{x})}\left[\log \frac{p(E_t, C_t \mid \mathbf{x})}{p(E_t \mid \mathbf{x})\, p(C_t \mid \mathbf{x})}\right]. \tag{4}$$

Intuitively, the CMI quantifies the reduction in uncertainty about a subject's survival status at time t if we know their censoring status, given that we have already accounted for the information in their covariates $\mathbf{X}$. The conditional independence hypothesis $E_t \perp C_t \mid \mathbf{X}$ holds if and only if $I(E_t; C_t \mid \mathbf{X}) = 0$. For discrete covariates $\mathbf{X}$, we can estimate the CMI empirically from a dataset of size n. The standard estimator is:

$$\hat{I}(E_t; C_t \mid \mathbf{X}) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{X}} \sum_{e_t \in \{0,1\}} \sum_{c_t \in \{0,1\}} N(e_t, c_t, \mathbf{x})\ \log\left(\frac{N(\mathbf{x})\ N(e_t, c_t, \mathbf{x})}{N(e_t, \mathbf{x})\ N(c_t, \mathbf{x})}\right), \tag{5}$$

where $N(\cdot)$ represents the empirical counts of observations. For instance, $N(e_t, c_t, \mathbf{x})$ is the number of subjects with covariates $\mathbf{x}$ having survival status e_t and censoring status c_t at time t, and $N(\mathbf{x})$ is the total number of subjects in that stratum. The stratum means the subgroup of the study population defined by a specific covariate profile $\mathbf{x}$. Formally, a stratum is the equivalence class of individuals sharing the same covariate vector $\mathbf{x}$.

4.1.1 CMI Bounds via Optimization over Feasible Completions

Due to right censoring, the empirical CMI in Equation 5 cannot be computed directly from the observed data. At a fixed time t, subjects censored before t have unknown event status E_t, while subjects with an event before t have unknown counterfactual censoring status C_t. Therefore, the contingency table of (E_t, C_t) is only partially observed.

We compute the CMI bounds by optimizing over all feasible completions of these unknown entries. For each covariate stratum $\mathbf{x}$, let $\ell_{\mathbf{x}}$ be the number of subjects censored before t, and let $m_{\mathbf{x}}$ be the number of subjects with an event before t. Define

$$u_{\mathbf{x}}^{10} \in \{0, \ldots, \ell_{\mathbf{x}}\}, \qquad u_{\mathbf{x}}^{01} \in \{0, \ldots, m_{\mathbf{x}}\},$$

where $u_{\mathbf{x}}^{10}$ is the number of censored subjects completed as $(E_t, C_t) = (1, 0)$, and $u_{\mathbf{x}}^{01}$ is the number of event subjects completed as $(E_t, C_t) = (0, 1)$. The remaining ambiguous subjects are completed as $(0, 0)$. Thus, the completed counts in stratum $\mathbf{x}$ are

$$\begin{aligned} N(1,1,\mathbf{x}) &= N^{\text{obs}}(1,1,\mathbf{x}), & N(1,0,\mathbf{x}) &= u_{\mathbf{x}}^{10}, \\ N(0,1,\mathbf{x}) &= u_{\mathbf{x}}^{01}, & N(0,0,\mathbf{x}) &= (\ell_{\mathbf{x}} - u_{\mathbf{x}}^{10}) + (m_{\mathbf{x}} - u_{\mathbf{x}}^{01}). \end{aligned} \tag{6}$$

Substituting these counts into Equation 5 gives the CMI associated with each feasible completion.

The CMI bounds at time t are then

$$I_{\min}^t = \min_{\{u_{\mathbf{x}}^{10}, u_{\mathbf{x}}^{01}\}_{\mathbf{x} \in \mathcal{X}}} \hat{I}(E_t; C_t \mid \mathbf{X}), \qquad I_{\max}^t = \max_{\{u_{\mathbf{x}}^{10}, u_{\mathbf{x}}^{01}\}_{\mathbf{x} \in \mathcal{X}}} \hat{I}(E_t; C_t \mid \mathbf{X}). \tag{7}$$

subject to

$$u_{\mathbf{x}}^{10} \in \{0, \ldots, \ell_{\mathbf{x}}\}, \qquad u_{\mathbf{x}}^{01} \in \{0, \ldots, m_{\mathbf{x}}\}, \qquad \forall \mathbf{x} \in \mathcal{X}.$$

The optimization is tractable because the objective decomposes across covariate strata. Hence, the global minimum or maximum can be obtained by solving independent two-variable integer programs for each stratum. In stratum $\mathbf{x}$, this requires enumerating only $(\ell_{\mathbf{x}}+1)(m_{\mathbf{x}}+1)$ possible completions, which avoids the exponential enumeration over all ambiguous subjects.

4.2 The MI Uncertainty Range and Permutation Testing

Because the full binary pair (E_t, C_t) is only partially observed, the empirical conditional mutual information $\hat{I}(E_t; C_t \mid \mathbf{X})$ is not uniquely determined from the observed data. Instead, each feasible completion of the unknown entries induces a possible value of the CMI. We therefore define the ***MI uncertainty range*** at time t as

$$\Delta I_t = I_{\max}^t - I_{\min}^t, \tag{8}$$

where

$$I_{\max}^t = \max_{\{u_{\mathbf{x}}^{10}, u_{\mathbf{x}}^{01}\}_{\mathbf{x} \in \mathcal{X}}} \hat{I}(E_t; C_t \mid \mathbf{X}), \qquad I_{\min}^t = \min_{\{u_{\mathbf{x}}^{10}, u_{\mathbf{x}}^{01}\}_{\mathbf{x} \in \mathcal{X}}} \hat{I}(E_t; C_t \mid \mathbf{X}). \tag{9}$$

Thus, ΔI_t measures the amount of uncertainty in the CMI that remains after accounting for all completions consistent with the observed censored data.

The proposed test does not assume that ΔI_t should be zero under independent censoring. Even when $E_t \perp C_t \mid X$, censoring can leave a nonzero range of feasible CMI values because the latent binary table is only partially observed. The relevant question is therefore whether the observed censored-data uncertainty is unusual relative to what would be expected under a conditionally independent event–censoring mechanism with comparable marginal behavior.

To answer this question, we construct a model-assisted permutation reference distribution. After estimating the marginal event and censoring processes, we generate null datasets by breaking the association between imputed event and censoring times within covariate strata while preserving the learned marginal distributions. For each null dataset, we recompute ΔI_t. The observed statistic is then compared with this reference distribution, and evidence is aggregated over landmark times to obtain a global diagnostic.

Why use uncertainty rather than one completion? The observed ΔI_t is unchanged by DVFM posterior draws when the data, strata, and landmarks are fixed. Moreover, $\Delta I_t/2$ is the smallest worst-case absolute error of representing all feasible empirical CMI values by a single number (Proposition 4). Thus the statistic measures unresolved dependence information rather than the dependence of one selected completed dataset. Proposition 5 also gives a two-candidate formula for its upper endpoint. These are analytical properties, not a claim that the final p-value is model-free or uniformly more powerful than alternative statistics.

5 A Model-Assisted Impute-and-Permute Test

We use $\Lambda^s = \max_{k=1,\ldots,K} \Delta I_{t_k}^s$ as the stratum statistic and $F = -2\sum_s \log p_s$ to aggregate its upper-tail ranks. Proposition 1 motivates searching across landmarks; power additionally requires separation of the uncertainty statistic itself. Fisher's expression is calibrated by reference ranks, not a chi-square distribution across strata.

We fit a joint DVFM and draw *one* compatible completed dataset $W_n = \{(\widetilde{E}_i, \widetilde{C}_i, \mathbf{X}_i)\}_{i=1}^n$. All B reference datasets reuse this completion: independently permute its censoring vector within exact covariate profiles, then reconstruct $T_i^{*(b)} = \min(\widetilde{E}_i, \widetilde{C}_{\pi_b(i)})$ and $\delta_i^{*(b)} = \mathbf{1}\{\widetilde{E}_i \le \widetilde{C}_{\pi_b(i)}\}$. The identity permutation reconstructs the original observations. Every dataset is processed by the same landmark rule and statistic map. Symmetric ranks include the original and all reference datasets on equal terms (Algorithms 1–2). Reusing the completion is part of the calibration construction, not treating imputed times as known truth. Its oracle validity and approximation bound are proved in Appendix F.3.

5.1 Deep Variational Frailty Model for Latent Imputation

5.1.1 Generative Architecture

DVFM specifies $p_\theta(e, c, z \mid \mathbf{x}) = f_{E,\theta}(e \mid \mathbf{x}, z) f_{C,\theta}(c \mid \mathbf{x}, z) p(z)$, with $p(z) = \mathcal{N}(0, I)$ and $E \perp C \mid (\mathbf{X}, Z)$. A decoder supplies positive Weibull parameters:

$$[k_E, \lambda_E, k_C, \lambda_C] = g_\theta(\mathbf{X}, z). \tag{10}$$

The shared latent permits dependence after marginalization; its survival covariance identity is given in Appendix F.1. The conditional Weibull functions are

$$S(t \mid k, \lambda) = \exp[-(t/\lambda)^k], \qquad f(t \mid k, \lambda) = \frac{k}{\lambda}(t/\lambda)^{k-1} \exp[-(t/\lambda)^k]. \tag{11}$$

5.1.2 Variational Inference and Conditional Completion

The encoder specifies $q_\phi(z \mid \mathbf{X}, T, \delta) = \mathcal{N}(\mu_\phi, \mathrm{diag}(\sigma_\phi^2))$. The training criterion is

$$\mathcal{L}_\beta(\theta, \phi; O_i) = \mathbb{E}_{q_\phi} \log L_\theta(O_i, z) - \beta D_{\mathrm{KL}}(q_\phi \| p), \tag{12}$$

where every margin is evaluated at $(T_i \mid \mathbf{X}_i, z)$ and

$$\log L_\theta(O_i, z) = \delta_i(\log f_E + \log S_C) + (1 - \delta_i)(\log S_E + \log f_C). \tag{13}$$

For $\beta = 1$, the exact ELBO identity is $\mathcal{L}_1 = \log p_\theta(T_i, \delta_i \mid \mathbf{X}_i) - D_{\mathrm{KL}}(q_\phi \| p_\theta(z \mid O_i))$. Other weights are regularized objectives, not automatically ordinary ELBOs. After fitting, draw $z_i \sim q_{\widehat{\phi}}(z \mid O_i)$ and the missing time from its normalized conditional distribution truncated above T_i, retaining the observed time. The likelihood derivation, an independent population-optimum result, and the complete sampling formulas appear in Appendices F.1–F.2. They justify the model and its completion operation, not unrestricted latent identification or automatic consistency of variational training.

5.2 Soundness and Power of the Test

We state the asymptotic guarantees of the proposed model-assisted impute-and-permute test. Let $\widehat{\Delta}_{s,k,n}$ denote the empirical MI-uncertainty width in covariate stratum s at landmark time t_k, and define the stratum-level statistic $\widehat{\Lambda}_{s,n} = \max_{k=1,\dots,K} \widehat{\Delta}_{s,k,n}$. Let $\widehat{\Lambda}^*_{s,n}$ denote the corresponding statistic computed from a null dataset generated by the DVFM-based impute-and-permute procedure.

Assumption 1 (Joint null-calibration consistency) *For each law P_0 satisfying $H_0 : E \perp C \mid \mathbf{X}$, let $\mathsf{P}_{n,B}$ be the joint law of the observed and B generated vectors of stratum maxima, including data, fitting, and resampling randomness. There exists a law $\mathsf{Q}_{n,B}$ that is exchangeable in the $B+1$ dataset indices and satisfies $d_{\mathrm{TV}}(\mathsf{P}_{n,B}, \mathsf{Q}_{n,B}) \le \varepsilon_{n,B}$. Along the considered sequence $B = B_n \to \infty$, $\varepsilon_{n,B_n} \to 0$. Here $d_{\mathrm{TV}}(P, Q) = \sup_A |P(A) - Q(A)|$. This is an explicit calibration assumption on the DVFM procedure, not a consequence of marginal consistency.*

Proposition 2 (Observed-law invariance of the MI-uncertainty statistic) *For every covariate stratum s and landmark time t, the population MI-uncertainty width ΔI_t^s is a functional only of the observed-data distribution of $(T, \delta, \mathbf{X})$. Consequently, if two latent event–censoring models induce the same observed-data distribution, then they induce the same population value of ΔI_t^s and the same sampling distribution of its empirical counterpart.*

Theorem 1 (Null validity under joint calibration) *Under Assumption 1, the symmetric ranks in Algorithm 2 satisfy, for $\alpha \in (0, 1)$,*

$$\Pr_{P_0}(\widehat{p}_{\mathrm{global},n,B} \le \alpha) \le \frac{\lfloor \alpha(B+1) \rfloor}{B+1} + \varepsilon_{n,B} \le \alpha + \varepsilon_{n,B}.$$

Hence $\limsup_{n\to\infty} \Pr_{P_0}(\widehat{p}_{\mathrm{global},n,B_n} \le \alpha) \le \alpha$, with finite-sample validity when $\varepsilon_{n,B} = 0$.

For the shared-completion construction, Theorem 3 allows $\varepsilon_{n,B} = r_n$, where r_n is the expected total-variation error in the conditional completion law. Correct conditional imputation gives finite-sample level control. Corollary 1 separates variational-posterior and fitted-model errors. These are sufficient conditions, not quantities certified by training loss alone.

Assumption 2 (Detectable upper-tail alternative) *Under $H_1 : E \not\perp C \mid \mathbf{X}$, let the number of strata J be fixed and $B_n \to \infty$. For some stratum $s^\star$, $\widehat{\Lambda}_{s^\star,n} \xrightarrow{p} \lambda_1$ while a reference maximum satisfies $\widehat{\Lambda}^*_{s^\star,n} \xrightarrow{p^*} \lambda_0$, where $\lambda_1 > \lambda_0$. The conditional convergence is in probability over the original data, fit, and shared completion; reference replicates are identically distributed conditional on them.*

Theorem 2 (Consistency against separated alternatives) *Under Assumption 2, the symmetric-rank procedure satisfies $\widehat{p}_{\mathrm{global},n,B_n} \xrightarrow{p} 0$. Consequently, for every fixed $\alpha \in (0,1)$,*

$$\Pr_{H_1}\{\widehat{p}_{\mathrm{global},n,B_n} \le \alpha\} \longrightarrow 1.$$

The separation concerns the actual maximum over the chosen landmarks; no independence between strata is required.

Remark 1 *The power result does not claim consistency against every possible distribution satisfying $E \not\perp C \mid \mathbf{X}$. Because the latent pair (E, C) is not fully identifiable from ordinary right-censored observations without additional assumptions, the theorem applies to identifiable alternatives that produce a detectable separation in the proposed statistic.*

6 Numerical Experiments on Synthetic Data

We evaluate the proposed test on synthetic data, where the ground truth is known, to assess its statistical power in detecting dependent censoring and its ability to control Type I error under independent censoring. We utilize four standard generative models based on copula functions: Gaussian (Li, 2000), Frank (Genest, 1987), Clayton (Clayton, 1978), and Gumbel (Gumbel, 1960). These models allow precise control over the dependence between the latent event time E and censoring time C, enabling evaluation across different dependence structures. To further assess the proposed method under dependence induced by a shared unobserved factor, we also consider a frailty-based data-generation model.

For each data-generating setting, with four features and 1000 instances, the proposed test was repeated over $R = 10$ random seeds with the number of quantiles $K \in \{3, 5, 7, 10\}$ and the number of permutations $B \in \{50, 100, 200, 500\}$. These repeated runs account for finite-sample variability, finite permutation noise, and stochastic components of the imputation procedure. Let $p_1, \ldots, p_R$ denote the resulting p-values. We summarize the repeated runs using two complementary quantities.

First, we combine evidence across repeated runs using Fisher's method,

$$X_F = -2\sum_{r=1}^{R} \log(p_r), \qquad p_{\mathrm{Fisher}} = \Pr\left(\chi^2_{2R} \ge X_F\right).$$

We also report the detection rate $\mathrm{DR}_{0.05} = \frac{1}{R}\sum_{r=1}^{R} \mathbf{1}\{p_r < 0.05\}$, which measures the stability of detection across random seeds. By Proposition 1, a statistically significant violation of the discrete-time null at any time point is sufficient to reject the continuous-time independent censoring assumption. Thus, detection does not require all repeated runs to produce p-values below 0.05. Instead, p_{Fisher} summarizes the accumulated evidence across repetitions, while $\mathrm{DR}_{0.05}$ quantifies robustness to finite-sample, permutation, and imputation variability. A small p_{Fisher} indicates evidence of dependent censoring, whereas a larger $\mathrm{DR}_{0.05}$ indicates that this evidence is more stable across random seeds.

For all synthetic experiments, the dependence parameter controls the strength of dependence between E and C. In the Clayton copula, $\theta = 0$ corresponds to independence and larger positive values of θ induce stronger lower-tail dependence. In the Frank copula, values close to zero correspond to weak dependence, while larger absolute values $|\theta|$ induce stronger dependence. In the Gaussian copula, $\theta = 0$ corresponds to independence, and larger $|\theta|$ indicates stronger positive or negative dependence. In the Gumbel copula, $\theta = 1$ corresponds to independence and larger values of θ induce stronger upper-tail dependence. For the frailty model, $\alpha = 0$ corresponds to independent censoring, while larger values of α create stronger dependence through the shared latent frailty variable.

The results in Tables 1 and 2 demonstrate that the proposed test controls false detection under independence and gains power as the dependence strength increases. In the independent settings,

Table 1: Repeated-run detection results for the four copula-based synthetic data-generation settings. The dependence parameter θ controls the strength of dependence: independence corresponds to $\theta = 0$ for Clayton and Gaussian copulas, $\theta = 1$ for the Gumbel copula, and values close to zero for the Frank copula. Bold and green entries indicate correct behavior: non-rejection under independence or rejection under dependence.

(a) Clayton copula

Parameter θ	p_{Fisher}	$\text{DR}_{0.05}$
0	**0.3596**	0.00
0.25	0.3907	0.10
0.8	0.2151	0.00
1	0.2525	0.10
2	0.9100	0.10
3	0.1770	0.10
5	0.2130	0.20
6	**0.0153**	0.40
10	**0.0224**	0.50
15	**0.0323**	0.60
20	**0.0371**	0.60
50	**0.0020**	0.80

(b) Frank copula

Parameter θ	p_{Fisher}	$\text{DR}_{0.05}$
-20	**0.0314**	0.30
-10	0.1256	0.10
-5	0.2648	0.10
-1	0.2639	0.00
0	**0.2794**	0.00
0.5	0.1921	0.10
1	**0.0086**	0.30
1.5	**0.0416**	0.50
3	**0.0163**	0.50
5	**0.0002**	0.60
10	**0.0251**	0.70
20	$\mathbf{9.92 \times 10^{-5}}$	0.70

(c) Gaussian copula

Parameter θ	p_{Fisher}	$\text{DR}_{0.05}$
-0.9	$\mathbf{6.28 \times 10^{-8}}$	0.60
-0.7	**0.0066**	0.50
-0.5	**0.0420**	0.50
-0.3	**0.0253**	0.30
0	**0.8001**	0.00
0.3	0.4306	0.10
0.5	0.4363	0.10
0.7	0.1028	0.00
0.9	**0.0430**	0.30

(d) Gumbel copula

Parameter θ	p_{Fisher}	$\text{DR}_{0.05}$
1	**0.6336**	0.00
1.4	0.1035	0.20
2	0.0843	0.00
3	0.0753	0.00
4	$\mathbf{9.61 \times 10^{-5}}$	0.50
5	$\mathbf{1.44 \times 10^{-9}}$	0.80
6	$\mathbf{6.86 \times 10^{-13}}$	1.00
10	$\mathbf{6.86 \times 10^{-13}}$	1.00
20	$\mathbf{6.86 \times 10^{-13}}$	1.00

Table 2: Repeated-run detection results for the frailty-based synthetic data-generation setting. The case $\alpha = 0$ corresponds to independent censoring, while larger values of α induce stronger dependence through the shared latent frailty variable. Bold entries indicate correct behavior: non-rejection under independence or rejection under dependence.

Parameter α	p_{Fisher}	$\text{DR}_{0.05}$
0	**0.9946**	0.00
1	0.4568	0.00
1.5	0.5266	0.00
1.7	**0.0323**	0.20
2	$\mathbf{7.69 \times 10^{-5}}$	0.20
3	$\mathbf{3.10 \times 10^{-10}}$	1.00
4	$\mathbf{1.70 \times 10^{-12}}$	1.00
5	$\mathbf{6.86 \times 10^{-13}}$	1.00
7	$\mathbf{6.86 \times 10^{-13}}$	1.00
10	$\mathbf{6.86 \times 10^{-13}}$	1.00
15	$\mathbf{6.86 \times 10^{-13}}$	1.00
20	$\mathbf{6.86 \times 10^{-13}}$	1.00

including Clayton $\theta = 0$, Gaussian and Frank $\theta = 0$, Gumbel $\theta = 1$, and frailty $\alpha = 0$, the Fisher combined p-values are large and $\text{DR}_{0.05} = 0$, indicating appropriate Type I error behavior (See Appendix H for more results). As the dependence parameter moves away from the independence

value, the Fisher combined p-values generally decrease and the detection rates increase, showing improved power under stronger dependent censoring.

The Fisher combined p-value and the detection rate provide complementary information. The value p_{Fisher} captures accumulated evidence across repeated runs, including cases where several runs produce moderately small p-values but not all are individually significant. In contrast, $\mathrm{DR}_{0.05}$ measures the stability of detection across runs. For example, in the frailty model with $\alpha = 2$, $\mathrm{DR}_{0.05} = 0.20$, but $p_{\mathrm{Fisher}} = 7.69 \times 10^{-5}$, indicating strong accumulated evidence despite variability across individual repetitions. For stronger frailty dependence, $\alpha \geq 3$, the detection becomes stable, with $\mathrm{DR}_{0.05} = 1.00$. Similar trends are observed for strong Gumbel dependence, where $\mathrm{DR}_{0.05}$ reaches 1.00 for $\theta \geq 6$.

The real-data results in Table 6 (Appendix E) indicate heterogeneous evidence of dependent censoring across datasets. AIDS and GBSG2 show no evidence against independent censoring, as both have large Fisher combined p-values and zero detection rate. In contrast, COVID, METABRIC, and NWTCO yield significant Fisher combined p-values, suggesting evidence of dependent censoring. The strongest and most robust signal is observed for NWTCO, where $p_{\mathrm{Fisher}} = 8.11 \times 10^{-10}$ and $\mathrm{DR}_{0.05} = 0.90$. METABRIC also shows clear evidence with moderate robustness, while COVID shows significant accumulated evidence but lower run-to-run stability. NACD and NPC have nonzero detection rates in a small fraction of runs, but their Fisher combined p-values are not significant, suggesting weaker or less stable evidence of dependent censoring. We compare against four full-likelihood copula diagnostics based on copula dependent-censoring models (Escarela and Carriere, 2003; Gharari et al., 2023) in Appendix C.1.

7 Conclusion

This paper introduced a principled framework for testing the independent censoring assumption in survival analysis, a central yet often unverifiable condition underlying many classical and modern time-to-event models. The key idea is to turn the ambiguity induced by censoring into a source of statistical information. To this end, we proposed the *MI uncertainty range* ΔI_t, defined as the width between the minimum and maximum feasible conditional mutual information values over all valid latent completions of the observed censored data. This statistic captures how the censored-data structure constrains the possible dependence between the latent event and censoring processes. To make the proposed statistic practical, we formulated the computation of the CMI bounds as a decomposable optimization problem over covariate strata, enabling exact and efficient evaluation without enumerating all global completions. We then embedded this statistic in a DVFM-based impute-and-permute testing procedure. The Deep Variational Frailty Model reconstructs latent event and censoring times through a shared frailty representation and Weibull marginal structure, while the stratified permutation step generates null datasets that enforce conditional independence while preserving the learned marginal behavior. Extensive experiments on synthetic datasets generated from copula-based and frailty-based mechanisms showed that the proposed test controls false detection under independent censoring and gains power as the dependence strength increases. Results on real datasets further demonstrated that the method can reveal heterogeneous evidence of possible dependent censoring across practical survival-analysis settings. Overall, the proposed framework provides a scalable and statistically motivated diagnostic tool for assessing independent censoring, thereby improving the reliability and interpretability of survival modeling under potentially informative censoring.

References

Norman E. Breslow and Nilanjan Chatterjee. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):457–468, 1999. doi: 10.1111/1467-9876.00165.

T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman. Survival analysis part i: basic concepts and first analyses. *British Journal of Cancer*, 89(2):232–238, 2003. doi: 10.1038/sj.bjc.6601118.

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. doi: 10.1093/biomet/65.1.141.

David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x.

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, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403): 346–352, 2012. doi: 10.1038/nature10983.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5 edition, 2019. doi: 10.1017/9781108591034.

Gabriel Escarela and Jacques F. Carriere. Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research*, 12(4):333–349, 2003. doi: 10.1191/0962280203sm335ra.

Philippe Flandre. Weighted log-rank test to compare two survival functions in the presence of dependent censoring. *Pharmaceutical Statistics*, 21(6):1281–1293, 2022. doi: 10.1002/pst.2245.

Christian Genest. Frank's family of bivariate distributions. *Biometrika*, 74(3):549–555, 1987. doi: 10.1093/biomet/74.3.549.

Ali Hossein Foomani Gharari, Michael Cooper, Russell Greiner, and Rahul G. Krishnan. Copula-based deep survival models for dependent censoring. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 669–680. PMLR, 2023. URL `https://proceedings.mlr.press/v216/gharari23a.html`.

Emil Julius Gumbel. Distributions des valeurs extrêmes en plusieurs dimensions. *Publications de l'Institut de Statistique de l'Université de Paris*, 9:171–173, 1960.

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. URL `https://www.jmlr.org/papers/v21/18-772.html`.

Scott M. Hammer, David A. Katzenstein, Michael D. Hughes, Holly Gundacker, Robert T. Schooley, Richard H. Haubrich, W. Keith Henry, Michael M. Lederman, John P. Phair, Manette Niu, Martin S. Hirsch, and Thomas C. Merigan. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine*, 335(15):1081–1090, 1996. doi: 10.1056/NEJM199610103351501.

Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *TEST*, 27(4):811–825, 2018. doi: 10.1007/s11749-017-0571-1.

Bingzhong Jing, Tao Zhang, Zixian Wang, Ying Jin, Kuiyuan Liu, Wenze Qiu, Liangru Ke, Ying Sun, Caisheng He, Dan Hou, Linquan Tang, Xing Lv, and Chaofeng Li. A deep survival analysis method based on ranking. *Artificial Intelligence in Medicine*, 98:1–9, 2019. doi: 10.1016/j.artmed.2019.06.001.

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. doi: 10.1186/s12874-018-0482-1.

Neeraj Kumar, Shi-ang Qi, Li-Hao Kuan, Weijie Sun, Jianfei Zhang, and Russell Greiner. Learning accurate personalized survival models for predicting hospital discharge and mortality of COVID-19 patients. *Scientific Reports*, 12:4472, 2022. doi: 10.1038/s41598-022-08601-6.

Seung-Yeoun Lee and Robert A. Wolfe. A simple test for independent censoring under the proportional hazards model. *Biometrics*, 54(3):1176–1182, sep 1998.

Seungyeoun Lee and Robert A. Wolfe. A class of linear nonparametric rank tests for dependent censoring. *Communications in Statistics—Theory and Methods*, 24(3):687–712, 1995. doi: 10.1080/03610929508831517.

David X. Li. On default correlation: A copula function approach. *Journal of Fixed Income*, 9(4): 43–54, 2000. doi: 10.3905/jfi.2000.319253.

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 of *Proceedings of Machine Learning Research*, pages 3841–3866. PMLR, 2026. URL https://proceedings.mlr.press/v337/lillelund26a.html.

Dimitris Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman and Hall/CRC, 2012. doi: 10.1201/b12208.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94 (448):1096–1120, 1999. doi: 10.1080/01621459.1999.10473862.

Martin Schumacher, G. Bastert, H. Bojar, K. Huebner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. L. Neumann, and H. F. Rauschecker. Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12(10):2086–2093, 1994. doi: 10.1200/JCO.1994.12.10.2086.

Yanqing Sun and Jimin Lee. Testing independent censoring for longitudinal data. *Statistica Sinica*, 21 (3):1315–1339, 2011. doi: 10.5705/ss.2009.251. URL https://www3.stat.sinica.edu.tw/sstest/oldpdf/A21n314.pdf.

Anastasios Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22, 1975. doi: 10.1073/pnas.72.1.20.

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, 2011.

A Appendix

Disclosure of AI Assistance

Generative AI tools were used to assist with manuscript drafting and editing, reference checking, and LaTeX formatting. The authors are responsible for the final scientific content and for the accuracy of this disclosure.

A.1 Limitations

The proposed test relies on some assumptions that affect its practical performance. First, the validity of the permutation null requires sufficiently populated covariate strata. When the number of strata is large relative to the sample size, the empirical CMI bounds may be noisy, reducing both calibration and power. Second, the DVFM-based imputation step assumes that the latent event and censoring marginals can be well approximated by the chosen Weibull decoder and latent frailty structure. If this structural model is severely misspecified, the reconstructed null distribution may be inaccurate.

Finally, the power guarantee applies to identifiable alternatives that induce a nonzero separation in the MI uncertainty statistic; some forms of dependent censoring may remain difficult to detect from right-censored data alone due to the Tsiatis nonidentifiability barrier.

B Proofs of the Theoretical Results

Throughout this appendix, $\mathbb{P}^*$ denotes probability conditional on the observed dataset, fitted DVFM, and the shared completed dataset used for permutation.

B.1 Proof of Proposition 1

Proof 1 *Suppose that*

$$E \perp C \mid \mathbf{X}.$$

Conditional independence is preserved under measurable transformations (Durrett, 2019). In particular, for any fixed $t > 0$, the mappings

$$e \mapsto \mathbf{1}\{e > t\}, \qquad c \mapsto \mathbf{1}\{c > t\}$$

are measurable. Hence,

$$\mathbf{1}\{E > t\} \perp \mathbf{1}\{C > t\} \mid \mathbf{X}.$$

Equivalently,

$$E_t \perp C_t \mid \mathbf{X}.$$

More explicitly,

$$\begin{aligned} P(E_t = 1, C_t = 1 \mid \mathbf{X}) &= P(E > t, C > t \mid \mathbf{X}) \\ &= P(E > t \mid \mathbf{X}) P(C > t \mid \mathbf{X}) \\ &= P(E_t = 1 \mid \mathbf{X}) P(C_t = 1 \mid \mathbf{X}). \end{aligned}$$

The factorization for the remaining values in $\{0,1\}^2$ follows by taking complements. Therefore, $E_t \perp C_t \mid \mathbf{X}$ for every $t > 0$.

The contrapositive gives

$$E_t \not\perp C_t \mid \mathbf{X} \text{ for some } t \quad \Longrightarrow \quad E \not\perp C \mid \mathbf{X}.$$

B.2 Proof of Proposition 2

Proof 2 *Fix a covariate stratum s and landmark time t. Every subject in this stratum belongs to one of the following three observable categories:*

$$\begin{aligned} A_t &= \{T > t\}, \\ D_t &= \{T \le t, \delta = 1\}, \\ R_t &= \{T \le t, \delta = 0\}. \end{aligned}$$

For subjects in A_t, the landmark pair is known to be

$$(E_t, C_t) = (1, 1).$$

For subjects in D_t, the event indicator is known to be zero and the censoring indicator is unknown, whereas for subjects in R_t, the censoring indicator is known to be zero and the event indicator is unknown.

Consequently, the complete feasible set of landmark contingency tables is determined entirely by the three observed counts

$$N_{A,s,t}, \qquad N_{D,s,t}, \qquad N_{R,s,t}.$$

The minimum and maximum feasible CMI values are therefore deterministic functions of these counts. Thus, for some deterministic function $\Psi_{s,t}$,

$$\widehat{\Delta I}_{s,t} = \Psi_{s,t}\left(N_{A,s,t}, N_{D,s,t}, N_{R,s,t}\right).$$

The joint distribution of these counts is determined entirely by the observed-data distribution of $(T, \delta, \mathbf{X})$. Hence, if two latent event–censoring models P and Q induce the same observed-data distribution,

$$P_{T,\delta,\mathbf{X}} = Q_{T,\delta,\mathbf{X}},$$

then the count vectors have the same distribution under P and Q. It follows that

$$\mathcal{L}_P\left(\widehat{\Delta I}_{s,t}\right) = \mathcal{L}_Q\left(\widehat{\Delta I}_{s,t}\right)$$

for every sample size.

Passing to the corresponding population probabilities shows that the population value $\Delta I_{s,t}$ is also the same under P and Q. Therefore, $\Delta I_{s,t}$ is a functional only of the observed-data distribution.

Remark 2 *This result does not require Tsiatis's nonidentifiability theorem. That theorem establishes that observationally equivalent independent and dependent latent models may exist. Proposition 2 establishes that whenever two such models have the same observed-data law, they necessarily have the same MI-uncertainty statistic.*

B.3 Proof of Theorem 1

Proof 3 *Fix P_0 under H_0, and set $m = B + 1$. Write $\Lambda_s^{(0)} = \Lambda_{\text{obs}}^s$ and $\Lambda_s^{(b)} = \Lambda_{\text{perm}}^{s,(b)}$ for $b \geq 1$. Algorithm 2 uses, for every $j = 0, \ldots, B$,*

$$p_s^{(j)} = \frac{1}{m}\sum_{\ell=0}^{B} \mathbf{1}\{\Lambda_s^{(\ell)} \geq \Lambda_s^{(j)}\}, \qquad F^{(j)} = -2\sum_s \log p_s^{(j)}.$$

Every rank includes itself exactly once, so $p_s^{(j)} \geq 1/m$. The observed and generated datasets are therefore ranked against the same collection of m statistic vectors.

The map from the joint array of stratum maxima to $(F^{(0)}, \ldots, F^{(B)})$ is permutation-equivariant in the dataset index: relabeling the datasets relabels their scores in the same way. Thus the score vector is exchangeable under $\mathsf{Q}_{n,B}$. For every j, define its upper rank

$$r_j = \frac{1}{m}\sum_{\ell=0}^{B} \mathbf{1}\{F^{(\ell)} \geq F^{(j)}\}.$$

The final-tail formula in Algorithm 2 is exactly $r_0 = \widehat{p}_{\text{global},n,B}$. For any deterministic score vector, at most $\lfloor m\alpha \rfloor$ indices have $r_j \leq \alpha$: ordering the scores from largest to smallest proves this count, and ties only increase upper ranks. Exchangeability therefore gives

$$\mathsf{Q}_{n,B}(r_0 \leq \alpha) = \mathbb{E}_{\mathsf{Q}_{n,B}}\left[\frac{1}{m}\sum_{j=0}^{B} \mathbf{1}\{r_j \leq \alpha\}\right] \leq \frac{\lfloor m\alpha \rfloor}{m}.$$

Since $\{r_0 \leq \alpha\}$ is a measurable event of the joint statistic array, Assumption 1 implies

$$\Pr_{P_0}(r_0 \leq \alpha) \leq \mathsf{Q}_{n,B}(r_0 \leq \alpha) + \varepsilon_{n,B} \leq \frac{\lfloor \alpha(B+1) \rfloor}{B+1} + \varepsilon_{n,B}.$$

This proves the finite-sample bound. Taking the upper limit along $B = B_n$ with $\varepsilon_{n,B_n} \to 0$ proves the asymptotic claim. The argument requires neither continuity of the score distribution nor independence between strata. It does not infer joint calibration from convergence of fitted marginal distributions.

B.4 Proof of Theorem 2

Proof 4 *Let $c = (\lambda_0 + \lambda_1)/2$ and suppress the informative stratum index. By assumption, $\Pr(\Lambda^{(0)} > c) \to 1$ and $q_n = \mathbb{P}^*(\Lambda^{(1)} \geq c) \to 0$ in probability. On the first event, the observed upper rank satisfies*

$$p^{(0)} \leq \frac{1}{B_n + 1} + \frac{1}{B_n + 1} \sum_{b=1}^{B_n} \mathbf{1}\{\Lambda^{(b)} \geq c\}.$$

The second term has conditional expectation at most q_n. Since $0 \leq q_n \leq 1$, its unconditional expectation tends to zero, and Markov's inequality gives $p^{(0)} \to 0$ in probability. Thus $F^{(0)} \geq -2 \log p^{(0)} \to \infty$.

Set $m = B_n + 1$. Within each stratum, its m upper ranks, in increasing order, are at least $1/m, \ldots, m/m$. Therefore, for every realized array,

$$\frac{1}{m} \sum_{j=0}^{B_n} F^{(j)} \leq \frac{2J}{m} \sum_{k=1}^{m} \log(m/k) \leq 2J,$$

where the last bound follows from $\log(m!) \geq \int_1^m \log x \, dx = m \log m - m + 1$. On $F^{(0)} > 0$, the global upper rank is at most $\{m^{-1} \sum_j F^{(j)}\}/F^{(0)} \leq 2J/F^{(0)}$, which converges to zero. This proves the assertion without an additional reference-tightness assumption or an interchange of Monte Carlo limits.

C Simulation of Survival Data Using Copulas

To study dependent censoring, we generate synthetic survival datasets where the joint dependence between latent event times E and censoring times C is introduced via copula functions. A copula $C_\theta(u, v)$ is a bivariate distribution function on $[0, 1]^2$ with uniform marginals, which allows us to couple any two marginal survival distributions into a valid joint distribution with a controlled dependence structure.

The simulation process involves two steps:

1. Sample a pair of dependent uniform random variables (U, V) from a chosen copula C_θ.
2. Transform these uniform variables into event and censoring times using their respective quantile functions (inverse transform sampling), i.e., $E = F_E^{-1}(U)$ and $C = F_C^{-1}(V)$.

We employ four standard parametric copula families. Below, we present the theoretical definition and the exact, well-established simulation algorithm used for each.

Gaussian Copula. The Gaussian copula is defined by the CDF of a bivariate standard normal distribution with correlation coefficient ρ.

$$C_\rho(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)), \quad \rho \in (-1, 1).$$

Simulation Algorithm: We generate a vector (Z_1, Z_2) from a bivariate standard normal distribution with correlation ρ (typically via Cholesky decomposition of the correlation matrix) and then set $U = \Phi(Z_1)$ and $V = \Phi(Z_2)$.

Clayton Copula. The Clayton copula is an Archimedean copula useful for modeling strong lower-tail dependence (i.e., early events are strongly associated with early censoring).

$$C_\theta(u, v) = \left[\max\{u^{-\theta} + v^{-\theta} - 1, 0\}\right]^{-1/\theta}, \quad \theta \in (0, \infty).$$

Simulation Algorithm (Marshall-Olkin): Generate a shared random variable $W \sim \mathrm{Gamma}(1/\theta, 1)$ and two independent variables $E_1, E_2 \sim \mathrm{Exp}(1)$. The dependent uniforms are then $U = (1 + E_1/W)^{-1/\theta}$ and $V = (1 + E_2/W)^{-1/\theta}$.

Frank Copula. The Frank copula is another Archimedean copula that allows for both positive and negative dependence. It is symmetric but has no asymptotic upper- or lower-tail dependence.

$$C_\theta(u,v) = -\frac{1}{\theta}\log\left(1 + \frac{(\exp(-\theta u)-1)(\exp(-\theta v)-1)}{\exp(-\theta)-1}\right), \quad \theta \in \mathbb{R}\setminus\{0\}.$$

Simulation Algorithm: For $\theta > 0$, let $p = 1 - \exp(-\theta)$. Generate $W \sim \mathrm{LogSeries}(p)$ and two independent variables $E_1, E_2 \sim \mathrm{Exp}(1)$. The dependent uniforms are

$$U = -\frac{1}{\theta}\log\left(1 - p\exp(-E_1/W)\right), \qquad V = -\frac{1}{\theta}\log\left(1 - p\exp(-E_2/W)\right).$$

For $\theta < 0$, first generate $(\widetilde{U}, \widetilde{V})$ from a Frank copula with parameter $|\theta|$, and then set $U = \widetilde{U}$ and $V = 1 - \widetilde{V}$. The limiting case $\theta = 0$ corresponds to independent uniform variables.

Gumbel Copula. The Gumbel copula is an Archimedean copula that exhibits upper-tail dependence.

$$C_\theta(u,v) = \exp\Big[-\left((-\log u)^\theta + (-\log v)^\theta\right)^{1/\theta}\Big], \quad \theta \in [1,\infty).$$

Simulation Algorithm: For $\theta > 1$, let $\alpha = 1/\theta$ and generate a positive α-stable random variable S with skewness 1, location 0, and scale

$$\left(\cos\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha} = \left(\cos\left(\frac{\pi}{2\theta}\right)\right)^{\theta}.$$

Then generate two independent variables $E_1, E_2 \sim \mathrm{Exp}(1)$. The dependent uniforms are

$$U = \exp\left[-\left(\frac{E_1}{S}\right)^{1/\theta}\right], \qquad V = \exp\left[-\left(\frac{E_2}{S}\right)^{1/\theta}\right].$$

The limiting case $\theta = 1$ corresponds to independent uniform variables. Our implementation relies on standard library functions for sampling the positive stable variable.

Fidelity of Simulation. While the theoretical CDFs are essential for mathematical analysis, the simulation algorithms we employ are the standard, exact methods for generating random variates from these distributions. For example, the Marshall-Olkin and Log-Series constructions are known to yield random variables with distributions that are mathematically identical to the Clayton and Frank copulas, respectively. This ensures that our synthetic data faithfully represents the desired dependence structures, providing a valid basis for evaluating our test's performance.

Data Generation via a Shared Frailty Model. This model induces dependence via an unobserved latent variable, or "frailty" $Z_i \sim \mathcal{N}(0,1)$, that simultaneously influences the event and censoring risks for each subject i. Given covariates $\mathbf{X}_i$, the proportional hazards for the event time E_i and censoring time C_i are:

$$\begin{aligned} h_E(t|\mathbf{X}_i, Z_i) &= h_{0,E}\exp(\,\mathbf{X}_i\boldsymbol{\beta} + \alpha_E Z_i\,) \\ h_C(t|\mathbf{X}_i, Z_i) &= h_{0,C}\exp(\,\mathbf{X}_i\boldsymbol{\gamma} + \alpha_C Z_i\,) \end{aligned}$$

The latent times E_i and C_i are sampled from exponential distributions with these rates, and the observed data are $T_i = \min(E_i, C_i)$ and $\delta_i = \mathbf{1}\{E_i \le C_i\}$. The censoring mechanism is controlled by the frailty coefficients: setting $\alpha_C = 0$ ensures conditionally independent censoring (H_0), as the frailty no longer links the two processes. Conversely, setting both $\alpha_E, \alpha_C \neq 0$ (typically with the same sign) induces dependent censoring (H_1), where the magnitude of the product $\alpha_E\alpha_C$ governs the dependence strength. The baseline hazard $_{0,C}$ is adjusted to achieve a target overall censoring rate.

C.1 Comparison with Full-Likelihood Copula Diagnostics

We compare the proposed diagnostic with four working-copula tests of conditional independent censoring. For each prespecified family $F \in \{\text{Clayton}, \text{Frank}, \text{Gaussian}, \text{Gumbel}\}$, the hypotheses are

$$H_0 : E \perp C \mid \mathbf{X} \qquad \text{against} \qquad H_{1,F} : E \not\perp C \mid \mathbf{X} \text{ within the working family } F.$$

Table 3: Full-likelihood parametric-bootstrap LRT p-values for four working survival-copula alternatives on the copula-generated datasets. A dagger marks the matched generating family under the numerical convention specified in Appendix C.1. Each entry is a separate test, using $n = 1000$, seed 42 and $B = 99$ bootstrap replicates; the minimum attainable p-value is 0.01. Bold entries satisfy the strict rule $p < 0.05$.

(a) Clayton-generated data

θ	Clayton†	Frank	Gaussian	Gumbel
0	1.00	0.28	0.56	1.00
0.25	0.23	0.26	0.34	0.47
0.8	**0.01**	0.16	0.32	1.00
1	**0.01**	0.19	0.23	1.00
2	**0.01**	0.11	0.10	0.36
3	**0.01**	0.18	0.22	0.30
5	**0.01**	0.08	0.14	0.14
6	**0.01**	o.10	**0.01**	0.07
10	**0.01**	**0.01**	**0.01**	0.05
15	**0.01**	**0.01**	**0.01**	**0.01**
20	**0.01**	**0.01**	**0.01**	**0.01**
50	**0.01**	**0.01**	**0.01**	**0.01**

(b) Frank-generated data

θ	Clayton	Frank†	Gaussian	Gumbel
-20	1.00	**0.01**	**0.01**	1.00
-10	1.00	**0.01**	**0.01**	1.00
-5	1.00	**0.02**	0.16	0.47
-1	0.23	0.32	0.66	1.00
-0.5	0.26	0.53	0.70	1.00
0.5	0.17	0.53	0.34	1.00
1	0.39	0.48	0.59	0.23
1.5	1.00	0.25	0.49	0.25
3	0.10	**0.04**	0.12	0.37
5	0.13	**0.03**	0.07	0.16
10	0.05	**0.01**	**0.01**	**0.03**
20	**0.01**	**0.01**	**0.01**	**0.01**

(c) Gaussian-generated data

θ	Clayton	Frank	Gaussian†	Gumbel
-0.9	1.00	**0.01**	**0.01**	1.00
-0.7	1.00	0.09	0.07	0.33
-0.5	1.00	0.15	0.15	1.00
-0.3	0.28	0.21	0.44	1.00
0	1.00	0.39	0.56	0.39
0.3	1.00	0.52	0.70	0.32
0.5	0.19	0.28	0.21	0.39
0.7	**0.02**	0.05	0.05	0.16
0.9	**0.01**	**0.03**	**0.01**	0.05

(d) Gumbel-generated data

θ	Clayton	Frank	Gaussian	Gumbel†
1	1.00	0.22	0.56	1.00
1.4	1.00	0.87	0.78	0.46
2	0.55	0.57	0.52	0.11
3	0.32	0.29	0.27	**0.01**
4	0.26	0.17	0.25	**0.01**
5	0.14	0.06	0.18	**0.01**
6	0.09	**0.01**	**0.04**	**0.01**
10	**0.01**	**0.01**	**0.01**	**0.01**
20	**0.01**	**0.01**	**0.01**	**0.01**

The null assumes independent Weibull proportional-hazards event and censoring margins. Under the alternative, both marginal parameter sets, their covariate coefficients, and the copula parameter are estimated jointly from $(\mathbf{X}, T, \delta)$. The likelihood-ratio statistic is

$$L_F = 2\{\ell_F(\widehat{\eta}_F, \widehat{\vartheta}_F) - \ell_0(\widehat{\eta}_0)\}.$$

Each p-value is computed by parametric bootstrap: sample independent event and censoring times from the fitted null, reconstruct the observed data, and refit both the null and the same working alternative. Let $L_F^{*(b)}$ denote the likelihood-ratio statistic recomputed in bootstrap replicate b. With B complete replicates,

$$p_F = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{L_F^{*(b)} \geq L_F\}}{B + 1}.$$

We use $B = 99$ complete bootstrap replicates in every setting. Values are consequently multiples of 0.01; $p_F = 0.01$ is the Monte Carlo resolution floor, not a claim that the underlying tail probability equals exactly 0.01. We use the strict rejection rule $p_F < 0.05$ throughout these tables.

Numerical copula convention. For these LRT experiments, the implementation uses a *survival copula*, $S_{E,C}(e, c \mid \mathbf{X}) = C_F\{S_E(e \mid \mathbf{X}), S_C(c \mid \mathbf{X})\}$. The associated generator therefore transforms sampled uniforms using $E = S_E^{-1}(U \mid \mathbf{X})$ and $C = S_C^{-1}(V \mid \mathbf{X})$. This specifies the inverse-survival convention of these numerical results, rather than the inverse-CDF convention described generically above. A dagger denotes a match under this convention. For the frailty generator, $\alpha_E = 1$ and $\alpha = \alpha_C$; all four families are working alternatives rather than exact generating copulas.

Table 4: Full-likelihood parametric-bootstrap LRT p-values on the shared-frailty generator. All four copulas are working alternatives. Results use $n = 1000$, seed 42, $B = 99$, and $\alpha_E = 1$, with the displayed $\alpha = \alpha_C$. Bold entries satisfy $p < 0.05$; a value of 0.05 is not counted as a rejection.

α	Clayton	Frank	Gaussian	Gumbel
0	0.01	0.69	0.02	0.16
1	0.01	0.27	0.16	0.23
1.5	0.01	0.27	0.12	0.22
1.7	0.01	0.44	0.20	0.10
2	0.01	0.68	0.15	0.53
3	0.01	0.17	0.19	0.23
4	0.01	0.35	**0.01**	0.32
5	0.01	0.10	**0.01**	**0.01**
7	0.03	**0.04**	**0.01**	**0.01**
10	0.01	**0.01**	**0.01**	**0.01**
15	0.01	**0.01**	**0.01**	**0.01**
20	0.01	**0.01**	**0.01**	**0.01**

Algorithm 1 One DVFM Conditional Completion

Require: Observed data D_n and fitted parameters $(\widehat{\theta}, \widehat{\phi})$.
Ensure: Compatible completion $W_n = \{(\widetilde{E}_i, \widetilde{C}_i, \mathbf{X}_i)\}_{i=1}^n$.

for $i = 1, \ldots, n$ **do**
  Draw $z_i \sim q_{\widehat{\phi}}(z \mid \mathbf{X}_i, T_i, \delta_i)$ and $U_i \sim \mathrm{Unif}(0, 1)$ independently.
  Decode $(k_{E,i}, \lambda_{E,i}, k_{C,i}, \lambda_{C,i})$ at $(\mathbf{X}_i, z_i)$.
  if $\delta_i = 1$ **then**
    $\widetilde{E}_i \leftarrow T_i$.
    $\widetilde{C}_i \leftarrow \lambda_{C,i}\{(T_i/\lambda_{C,i})^{k_{C,i}} - \log U_i\}^{1/k_{C,i}}$.
  else
    $\widetilde{C}_i \leftarrow T_i$.
    $\widetilde{E}_i \leftarrow \lambda_{E,i}\{(T_i/\lambda_{E,i})^{k_{E,i}} - \log U_i\}^{1/k_{E,i}}$.
  end if
end for
return W_n. ▷ Reuse this completion for the entire permutation test.

Algorithm 2 MI-Uncertainty Test with Shared-Completion Calibration

Require: Data D_n, B, exact covariate strata, and a prespecified landmark rule $\mathcal{T}(\cdot)$.

Fit DVFM once to D_n; draw W_n once using Algorithm 1.
$D_n^{(0)} \leftarrow D_n$.
for $b = 1, \ldots, B$ **do**
    Draw π_b uniformly from all within-stratum permutations, independently across b.
    Retain $\widetilde{E}_i$ and replace $\widetilde{C}_i$ by $\widetilde{C}_{\pi_b(i)}$.
    Form $D_n^{(b)} = \{(\min(\widetilde{E}_i, \widetilde{C}_{\pi_b(i)}), \mathbf{1}\{\widetilde{E}_i \le \widetilde{C}_{\pi_b(i)}\}, \mathbf{X}_i)\}_{i=1}^n$.
end for
for $j = 0, \ldots, B$ **do**
    Apply the same rule $\mathcal{T}(D_n^{(j)})$ and compute $\Lambda_s^{(j)} = \max_k \widehat{\Delta}_{s,k}^{(j)}$ in every stratum.
end for
for $j = 0, \ldots, B$ **do**
    for all stratum s **do**
        $p_s^{(j)} \leftarrow (B+1)^{-1} \sum_{\ell=0}^{B} \mathbf{1}\{\Lambda_s^{(\ell)} \ge \Lambda_s^{(j)}\}$.
    end for
    $F^{(j)} \leftarrow -2 \sum_s \log p_s^{(j)}$.
end for
$p_{\text{global}} \leftarrow (B+1)^{-1} \sum_{j=0}^{B} \mathbf{1}\{F^{(j)} \ge F^{(0)}\}$.
return p_{global}.

D ILLUSTRATION OF THE LANDMARK REPRESENTATION

Table 5: Discretization of continuous survival data. The observed data (T, δ) generate a sequence of discrete-time survival (E_t) and censoring (C_t) indicators, which depend on the underlying latent continuous times (E, C). "?" denotes values that are unobserved due to a prior event or censoring.

	Observed Data		**Latent Times**		$t=0$		$t=1.5$		$t=2.1$		$t=3.0$		$t=5.0$	
Sub.	T	δ	**E**	**C**	E_0	C_0	$E_{1.5}$	$C_{1.5}$	$E_{2.1}$	$C_{2.1}$	E_3	C_3	E_5	C_5
A	1.5	1	1.5	≥ 1.5	1	1	**0**	?	**0**	?	**0**	?	**0**	?
B	2.1	1	2.1	≥ 2.1	1	1	1	1	**0**	?	**0**	?	**0**	?
C	3.0	0	> 3.0	3.0	1	1	1	1	1	1	?	**0**	?	**0**
D	5.0	0	> 5.0	5.0	1	1	1	1	1	1	1	1	?	**0**

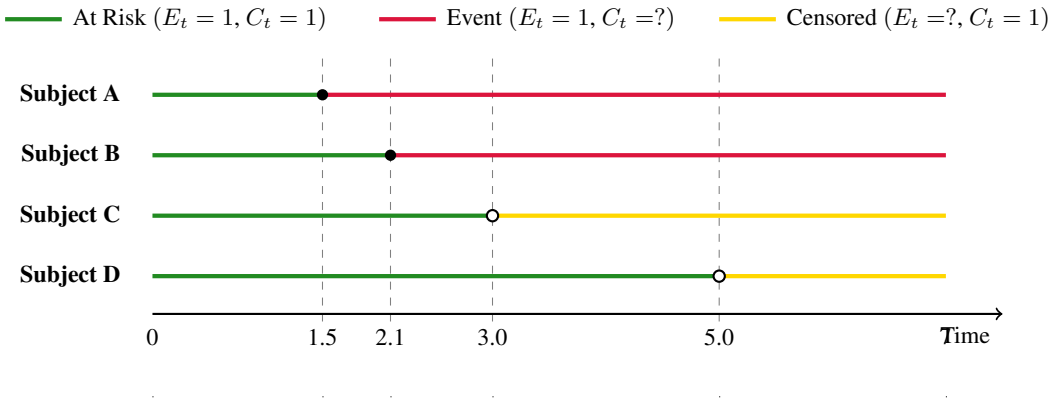

Figure 1: Timeline visualization of survival data for four subjects. A single line represents each subject's status at each $t > 0$: 'At Risk' (green: $E_t = 1$, $C_t = 1$), 'Event Occurred earlier' (red: $E_t = 0$. $C_t =?$), or 'Censored earlier' (yellow: $E_t =?$, $C_t = 0$). An event ($\delta = 1$) is marked with a filled circle, and censoring ($\delta = 0$) is marked with an open circle.

E REAL-DATA NUMERICAL SUMMARY

Table 6: Repeated-run detection results on real datasets. For each dataset, the proposed test was repeated over 10 runs.

Dataset	# Samples	# Covariates	# Strata	p_{Fisher}	$\mathrm{DR}_{0.05}$
AIDS (Hammer et al., 1996)	2139	6–8	1–6	0.3917	0.00
COVID (Kumar et al., 2022)	1422	6	4–12	0.0058	0.20
GBSG2 (Schumacher et al., 1994)	686	4–6	1–3	0.9394	0.00
METABRIC (Curtis et al., 2012)	1981	6	3–50	0.0008	0.50
NACD (Haider et al., 2020)	2396	7–9	2–15	0.3435	0.10
NPC (Jing et al., 2019)	6449	7–8	3–48	0.1108	0.10
NWTCO (Breslow and Chatterjee, 1999)	4028	3	9–11	8.11×10^{-10}	0.90

F SELF-CONTAINED FOUNDATIONS FOR IMPUTATION AND CALIBRATION

We derive the conditional completion law and show how its accuracy enters the Type-I error bound for Algorithms 1–2. The argument separates model representation, variational approximation, and permutation calibration. It concerns distributions of missing times rather than identification of neural parameters or recovery of individual latent coordinates.

F.1 JOINT MODEL, OBSERVED LIKELIHOOD, AND VARIATIONAL OBJECTIVE

Let $o = (t, \delta, \mathbf{x})$, and let Z have prior $p(z) = \mathcal{N}(0, I)$. Conditional on $(\mathbf{X}, Z)$, assume independent, positive, continuous event and censoring times. The joint model is

$$p_\theta(e, c, z \mid \mathbf{x}) = f_{E,\theta}(e \mid \mathbf{x}, z) f_{C,\theta}(c \mid \mathbf{x}, z) p(z).$$

The decoder supplies positive Weibull shapes and scales to both margins. The encoder supplies the mean and diagonal covariance of $q_\phi(z \mid o)$. The model-implied observed density is

$$L_\theta(o, z) = \{f_{E,\theta}(t \mid \mathbf{x}, z) S_{C,\theta}(t \mid \mathbf{x}, z)\}^\delta \{S_{E,\theta}(t \mid \mathbf{x}, z) f_{C,\theta}(t \mid \mathbf{x}, z)\}^{1-\delta}, \tag{14}$$

$$p_\theta(t, \delta \mid \mathbf{x}) = \int L_\theta(o, z) p(z)\, dz, \qquad p_\theta(z \mid o) = \frac{L_\theta(o, z) p(z)}{p_\theta(t, \delta \mid \mathbf{x})}. \tag{15}$$

For $\delta = 1$, the observed-density contribution follows by integrating $f_E(t \mid \mathbf{x}, z) f_C(c \mid \mathbf{x}, z)$ over $c > t$; the other contribution is obtained by exchanging E and C. The joint survival function and its departure from independence satisfy

$$\begin{aligned} S_\theta(e, c \mid \mathbf{x}) &= \mathbb{E}_{p(z)}[S_{E,\theta}(e \mid \mathbf{x}, Z) S_{C,\theta}(c \mid \mathbf{x}, Z)], \\ S_\theta(e, c \mid \mathbf{x}) - S_{E,\theta}(e \mid \mathbf{x}) S_{C,\theta}(c \mid \mathbf{x}) &= \mathrm{Cov}_{p(z)}\{S_{E,\theta}(e \mid \mathbf{x}, Z), S_{C,\theta}(c \mid \mathbf{x}, Z)\}. \end{aligned} \tag{16}$$

This is just the definition of covariance after marginalizing the latent. It explains the dependence mechanism; it does not identify that mechanism from the observed likelihood.

For the unweighted KL objective,

$$\begin{aligned} \mathcal{L}_1(\theta, \phi; o) &= \mathbb{E}_{q_\phi(z \mid o)} \log L_\theta(o, z) - D_{\mathrm{KL}}\{q_\phi(z \mid o) \| p(z)\} \\ &= \log p_\theta(t, \delta \mid \mathbf{x}) - D_{\mathrm{KL}}\{q_\phi(z \mid o) \| p_\theta(z \mid o)\}. \end{aligned} \tag{17}$$

The identity follows by substituting the posterior in Equation 15 into the KL divergence. It proves the ordinary evidence lower bound. For a KL weight β,

$$\mathcal{L}_\beta = \mathcal{L}_1 + (1 - \beta) D_{\mathrm{KL}}(q_\phi \| p).$$

Thus $\beta \neq 1$ changes the objective: for $\beta \geq 1$ it is still a lower bound, but for $0 \leq \beta < 1$ a lower-bound interpretation is not guaranteed. Shrinkage penalties and warm-up schedules are optimization choices. The identities require normalized distributions and positive decoder parameters, but do not establish convergence of a training algorithm.

Proposition 3 (Compatibility with an independent population optimum) *Suppose $E \perp C \mid \mathbf{X}$ under P_0, and the decoder contains parameters θ_0 for which, for $j \in \{E, C\}$,*

$$f_{j,\theta_0}(u \mid \mathbf{x}, z) = f_{j,0}(u \mid \mathbf{x})$$

for $P_{0,\mathbf{X}}$-almost every $\mathbf{x}$ and p-almost every z. Assume the variational family contains $q_{\phi_0}(z \mid o) = p(z)$ and the expectations below are finite. Then (θ_0, ϕ_0) maximizes the population ordinary ELBO. Its exact conditional completion law is the true null completion law.

Proof 5 *The KL identity and conditional cross-entropy inequality give*

$$\mathbb{E}_0 \mathcal{L}_1(\theta, \phi; O) \leq \mathbb{E}_0 \log p_\theta(T, \delta \mid \mathbf{X}) \leq \mathbb{E}_0 \log p_0(T, \delta \mid \mathbf{X}).$$

At θ_0, the observed likelihood is $p_0(t, \delta \mid \mathbf{x})$ and does not depend on z, so the exact latent posterior is $p(z)$. Choosing $q_{\phi_0} = p$ attains both equalities. Both fitted margins equal the true margins, so their truncated conditional laws are also correct.

This is an existence statement under correct marginal specification, not uniqueness or a guarantee about numerical optimization. In particular, it requires the *two margins*, not merely the observed density, to be represented correctly. If two candidates induce the same observed density and both have exact variational posteriors, their ordinary population ELBOs tie. Observed-likelihood fit alone therefore cannot certify the imputation law. Conversely, the existence of this null-compatible submodel shows that shared-latent modeling does not inherently force dependence.

F.2 Posterior completion, rather than unconditional latent prediction

For $O_i = (T_i, \delta_i, \mathbf{X}_i)$, retain the observed time and sample $z_i \sim q_{\widehat{\phi}}(z \mid O_i)$. Given this draw, if $\delta_i = 1$, set $\widetilde{E}_i = T_i$ and sample the missing time from

$$\widehat{f}_C(c \mid \mathbf{X}_i, z_i, C > T_i) = \frac{f_{C,\widehat{\theta}}(c \mid \mathbf{X}_i, z_i)\mathbf{1}\{c > T_i\}}{S_{C,\widehat{\theta}}(T_i \mid \mathbf{X}_i, z_i)}. \tag{18}$$

For $\delta_i = 0$, set $\widetilde{C}_i = T_i$ and use the analogous event law truncated to $e > T_i$. These distributions are normalized and keep every completion compatible with the observed data. For a Weibull missing margin, a direct draw is

$$U_i \sim \mathrm{Unif}(0,1), \qquad T_{\mathrm{mis},i} = \lambda_i \left\{ (T_i/\lambda_i)^{k_i} - \log U_i \right\}^{1/k_i}. \tag{19}$$

Indeed, the Weibull cumulative hazard is exponential and its residual above $(T_i/\lambda_i)^{k_i}$ is again exponential. Here (k_i, λ_i) are those of the *missing* margin evaluated at $(\mathbf{X}_i, z_i)$.

Let $K^{\mathrm{ex}}_{\theta,i}(\cdot \mid O_i)$ denote the law obtained by using the exact posterior $p_\theta(z \mid O_i)$ instead of q_ϕ. By conditioning first on z and then integrating it out, this is precisely the model's conditional law of the missing time given O_i. An aggregate posterior over the cohort is a different distribution and is not substituted for this subject-specific posterior. Let $K_{0,n}(\cdot \mid D_n)$ denote the true conditional law of all missing times given the observed dataset.

Lemma 1 (Oracle completion identity) *For any joint data-generating law with a regular conditional completion kernel, draw D_n from its observed-data law and then draw a compatible completed dataset W_n from $K_{0,n}(\cdot \mid D_n)$. The marginal law of W_n is the true complete-data law, and coarsening W_n returns D_n almost surely.*

Proof 6 *For every measurable complete-data set A, the tower property gives $\Pr(W_n \in A) = \mathbb{E}[K_{0,n}(A \mid D_n)] = \Pr(W_n^{\mathrm{true}} \in A)$. The equality after coarsening follows from the support of the conditional kernel. No independent-censoring assumption is needed for this identity.*

F.3 A constructive calibration bound for a shared completion

The following argument uses *one completed dataset per test execution*. All B reference datasets are obtained by independently permuting the censoring vector of that same completion. The original observed dataset is the coarsening of its identity permutation. Re-imputing separately inside every permutation loop is a different joint construction and is not covered by this argument.

Condition on the observed covariate design. Subjects are independent, and the permutations are within exact covariate profiles; under H_0, subjects within a stratum have the same event law and the same censoring law. More general strata are admissible only if this within-stratum invariance is justified. Let $\mathcal{G}$ be the finite group that permutes censoring times within those strata and leaves event times fixed. Let g_0 be the identity and $g_1, \ldots, g_B$ be independent uniform draws from the *whole* group, independently of the completion.

The statistic map applied to each coarsened dataset must be the same. Landmarks are either fixed independently of the outcomes or selected by the same prespecified measurable rule separately on each coarsened dataset. An observed-only, data-selected grid is not frozen for the reference datasets unless its permutation invariance is separately established. Hyperparameters and external random choices used by this map are fixed beforehand or conditioned on independently of the outcomes. The MI bound computation itself is deterministic once the observed dataset and its landmark rule are fixed.

Write $\widehat{K}_n(\cdot \mid D_n)$ for the DVFM completion kernel conditional on the fitted model, and set

$$r_n = \mathbb{E}_0\Big[d_{\mathrm{TV}}\{\widehat{K}_n(\cdot \mid D_n), K_{0,n}(\cdot \mid D_n)\} \Big], \qquad d_{\mathrm{TV}}(P,Q) = \sup_A |P(A) - Q(A)|. \tag{20}$$

The expectation includes data and fitting randomness; conditioning on the covariate design is understood. The same result holds after averaging over that design.

Theorem 3 (Calibration through conditional completion accuracy) *Under H_0 and the construction above, the joint statistic law is within r_n in total variation of an exchangeable law, for every $B \geq 1$. Consequently, the symmetric-rank p-value satisfies*

$$\Pr_0\{\widehat{p}_{\mathrm{global},n,B} \leq \alpha\} \leq \frac{\lfloor \alpha(B+1) \rfloor}{B+1} + r_n \leq \alpha + r_n, \qquad 0 < \alpha < 1. \tag{21}$$

Exact conditional completion gives finite-sample level control. If $r_n \to 0$, asymptotic level control holds along any chosen sequence B_n; there is no additional factor B_n multiplying r_n.

Proof 7 *First use $K_{0,n}$. By Lemma 1, the completed data have the true complete-data null law, which is invariant under $\mathcal{G}$. Conditional on their finite permutation orbit, they are uniform on that orbit. Uniform independent group elements send any orbit member to a uniform orbit member. Hence $W_n, g_1 W_n, \ldots, g_B W_n$ are conditionally i.i.d. on the orbit, and therefore exchangeable. Applying the common coarsening and statistic maps preserves exchangeability. This is the group-invariance basis of random permutation inference (Hemerik and Goeman, 2018).*

Replacing $K_{0,n}$ by $\widehat{K}_n$ changes the marginal completion law by at most r_n: integrate the conditional total-variation bound over D_n and fitting randomness. Every subsequent operation, including all B permutations, is the same Markov kernel applied to that one completion, so total variation cannot increase. Thus the full statistic-array laws are within r_n. The rank-counting proof of Theorem 1 then gives Equation 21. Since only one completion kernel is perturbed, the bound contains no factor B.

What this establishes. Theorem 3 supplies an explicit sufficient route to Assumption 1: take $\varepsilon_{n,B} = r_n$. It separates the exact permutation argument from estimation and variational errors. The true completion law is not available in an application, so r_n is not a data-certified error bound. Empirical calibration of the fitted procedure is assessed separately from prediction, joint-survival recovery, and observed-likelihood fit.

F.4 Separating variational and fitted-model error

Conditional on the fitted parameters, use independent posterior draws and independent truncated draws across subjects. Define

$$\begin{aligned} G_i &= D_{\mathrm{KL}}\{q_{\widehat{\phi}}(z \mid O_i) \| p_{\widehat{\theta}}(z \mid O_i)\}, \\ b_i &= d_{\mathrm{TV}}\{K^{\mathrm{ex}}_{\widehat{\theta},i}(\cdot \mid O_i), K_{0,i}(\cdot \mid O_i)\}. \end{aligned} \tag{22}$$

The first measures the posterior approximation gap *within* the fitted model. The second includes model misspecification and estimation error in the actual conditional missing-time law.

Corollary 1 (Completion-error decomposition) *With the preceding definitions,*

$$r_n \leq \mathbb{E}_0 \min\left\{1, \sqrt{\frac{1}{2}\sum_{i=1}^{n} G_i} + \sum_{i=1}^{n} b_i\right\}. \tag{23}$$

If P_0 is correctly represented and both the exact fitted completion laws and the variational posteriors are correct, the right-hand side is zero.

Proof 8 *Introduce the intermediate product of exact fitted completion kernels. The KL divergence between the product variational latent posteriors and product exact fitted latent posteriors is $\sum_i G_i$. Pinsker's inequality bounds their total variation by $\sqrt{\sum_i G_i/2}$. Applying the same conditional truncated-time sampling kernels cannot increase this distance. The distance between the product exact fitted completion laws and the true product completion laws is at most $\sum_i b_i$, by the product coupling bound. Apply the triangle inequality, cap the distance at one, and average.*

The sum-based bound is sufficient and can be conservative; it is not a necessary rate condition for the statistic. In particular, small average training loss does not show that either term vanishes. It does identify *which* approximations matter, while allowing arbitrary nonlinear decoders and arbitrary dependence between the final stratum scores.

G Analytical Role of the MI-Uncertainty Statistic

The following are properties of the observed-data statistic, not claims of empirical power superiority. They distinguish representation of uncertainty from prediction of a particular completed dataset.

G.1 Completion invariance and unavoidable point-estimation ambiguity

Fix an observed dataset D and its covariates and landmark. Let $\mathcal{Q}_t(D)$ be the finite set of feasible empirical completed tables, and let $I(Q)$ be their empirical CMI. Write $L_t = \min_Q I(Q)$, $U_t = \max_Q I(Q)$, and $\Delta I_t = U_t - L_t$.

Proposition 4 (Sharp ambiguity and minimax interpretation) *The interval $[L_t, U_t]$ is the smallest closed interval containing the CMI of every feasible completion, and*

$$\inf_{a\in\mathbb{R}} \max_{Q\in\mathcal{Q}_t(D)} |a - I(Q)| = \frac{\Delta I_t}{2}. \tag{24}$$

The midpoint $(L_t + U_t)/2$ attains the infimum. The width and the observed maximum over any fixed collection of landmarks are unchanged by all DVFM posterior draws or parameterizations when D, its strata, and its landmarks are held fixed.

Proof 9 *Both endpoints are attained because the feasible set is finite. Any interval containing all feasible CMI values must contain the endpoints. For any a,*

$$\max_Q |a - I(Q)| \geq \max\{|a - L_t|, |a - U_t|\} \geq (U_t - L_t)/2.$$

The midpoint has distance at most $(U_t - L_t)/2$ from every value between the endpoints. Finally, the feasible tables depend only on the observed counts, not on the DVFM or a sampled completion, which proves the invariance.

Thus $\Delta I_t/2$ measures the smallest worst-case error of representing all censoring-compatible empirical CMI values by one number. A completed-data MI statistic chooses one feasible table (or an imputation-weighted average); the uncertainty statistic instead reports how much remains undetermined by the observations. This is an exact finite-sample statement about the feasible *empirical* CMI values, not a confidence interval for population CMI and not a claim of robustness of the final p-value to a wrong null generator. The absence of completion-draw variability concerns the observed statistic only; null calibration and data-selected landmarks remain sources of variation.

G.2 An exact expression for the upper endpoint

Fix a stratum and a landmark with no observed time equal to it. Let a, d, r be the fractions with $T > t$, with $T < t, \delta = 1$, and with $T < t, \delta = 0$, respectively. Thus $a + d + r = 1$. Every normalized feasible table has, in the order $(00, 01, 10, 11)$, the form

$$Q(x, y) = (1 - a - x - y,\ x,\ y,\ a), \qquad 0 \leq x \leq d, \quad 0 \leq y \leq r.$$

For empirical tables, x and y range over the corresponding integer-count grid. Define $h(u) = -u \log u - (1 - u)\log(1 - u)$ and $H(a, d, r) = -a \log a - d \log d - r \log r$, with $0 \log 0 = 0$.

Proposition 5 (Two-candidate maximum) *For both the continuous feasible rectangle and its empirical integer grid,*

$$I^t_{\max} = \max\{h(a),\ h(d) + h(r) - H(a, d, r)\}. \tag{25}$$

For the population rectangle, $I^t_{\min} = 0$. For empirical tables, $I^t_{\min}$ may be positive because of integer restrictions; it remains computed by the stated exact finite optimization.

Proof 10 *For fixed y, the event marginal is fixed and MI is convex in the conditional censoring probabilities, hence convex in x. More explicitly, writing $q_{00} = 1 - a - x - y$, its interior second derivative is*

$$\frac{\partial^2 I}{\partial x^2} = \frac{1}{x} - \frac{1}{x + a} + \frac{1}{q_{00}} - \frac{1}{q_{00} + y} \geq 0.$$

The same argument with x, y interchanged gives convexity in y; boundary cases follow by continuity. The maximum is therefore attained at a rectangle corner, which also lies in the empirical grid. The $(0, 0)$ corner has MI $h(a)$. At each mixed corner, one binary marginal has entropy $h(a)$, so MI cannot exceed $h(a)$. At (d, r), $q_{00} = 0$ and MI equals $h(d) + h(r) - H(a, d, r)$. This proves the maximum formula.

For the population minimum, if $a + r > 0$, set $y = r$ and $x = ad/(a + r)$. The resulting table

$$\left(\frac{dr}{a+r}, \frac{ad}{a+r}, r, a\right)$$

factorizes and therefore has MI zero. If $a + r = 0$, the event indicator is constant and MI is also zero. Nonnegativity proves the minimum claim.

This reduces *maximization* to two entropy evaluations; it does not remove the empirical minimization problem or establish a speed-up of the whole neural fitting procedure. It also makes the statistic's scope clear: when $a = 1/2$, $I^t_{\max} = \log 2$ and the population width is saturated. A median landmark can consequently be uninformative about an upper-tail separation. The actual maximum must satisfy the separation assumption of Theorem 2; positive dependence alone is not that assumption.

What remains for empirical comparison. The analytical advantages established here are completion-invariant observed measurement, an exact interpretation of unresolved MI, and a closed-form upper endpoint. Whether they improve rejection rates over another statistic with the same fitted DVFM and permutation construction is an empirical question. Such a matched-statistic comparison and null-calibration experiments must evaluate the complete, shared-completion procedure; prediction metrics from a different training protocol cannot replace them.

H Null Calibration under Independent Censoring

H.1 Experimental design

We examine the finite-sample Type-I error of the proposed MI-uncertainty test when $E \perp C \mid \mathbf{X}$. The study comprises 60 independently generated datasets, each containing $n = 1000$ subjects and four independent Bernoulli$(1/2)$ covariates. We generate $R = 10$ datasets at each of the four copula independence limits: Clayton $\theta = 0$, Frank $\theta = 0$, Gaussian $\rho = 0$, and Gumbel $\theta = 1$. At independence, these labels correspond to separate simulation batches with the same family of independent margins. We also consider two frailty nulls, $(\alpha_E, \alpha_C) = (0, 0)$ and $(1, 0)$, with ten datasets each. The latter introduces latent heterogeneity in the event process without inducing event–censoring dependence.

For each replication, we independently draw the coefficient vectors $\beta_E, \beta_C \sim \mathrm{Unif}([-0.5, 0.5]^4)$ and keep them fixed within that dataset. For the copula-null batches, independent uniforms U, V give

$$E = 100\{-\log U / \exp(\mathbf{X}^\top \beta_E)\}^{1/1.5}, \qquad C = 150\{-\log V / \exp(\mathbf{X}^\top \beta_C)\}.$$

For the frailty nulls, we draw $Z \sim \mathcal{N}(0, 1)$ and use

$$E \mid \mathbf{X}, Z \sim \mathrm{Exp}\{\exp(\mathbf{X}^\top \beta_E + \alpha_E Z)\}, \qquad C \mid \mathbf{X} \sim \mathrm{Exp}\{b_C \exp(\mathbf{X}^\top \beta_C)\},$$

where the exponential parameters are rates and $\alpha_C = 0$ in both designs. The censoring scale b_C is determined before sampling evaluation times by a population-expectation calculation targeting 30% censoring. This calculation averages over all 16 binary covariate profiles and uses 100-point Gaussian quadrature for the event-only frailty model. The observed data are $T = \min(E, C)$ and $\delta = \mathbf{1}\{E \le C\}$. Only $(\mathbf{X}, T, \delta)$ are supplied to the learned detector.

H.2 Fitting and testing protocol

The DVFM uses encoder widths $(64, 32)$, decoder widths $(32, 64)$, ReLU activations, batch normalization, and latent dimension $d_z = 1000$. Training uses Adam with learning rate 10^{-3}, batch size 64, and 200 epochs; the KL weight increases to one over the first 50 epochs. Times are normalized by their observed sample maximum. Each dataset has a separate generation and fitting seed, with root seed 20260925 defining independent streams for data generation, hyperparameter selection, fitting, conditional completion, and permutation. All 60 fits completed.

For each replication, K and B are sampled from $\{3, 5, 7, 10\}$ and $\{50, 100, 200, 500\}$, respectively, before examining the test outputs. Strata are defined by exact four-covariate profiles, with a minimum size of 20; all 16 profiles satisfy this criterion in every dataset. We draw one censoring-compatible posterior completion and reuse it for all B independent within-stratum permutations, as in Algorithms 1–2.

Landmarks are linearly interpolated empirical quantiles with probabilities evenly spaced from 0.15 to 0.85. The same selection rule is applied separately to the original dataset and each reconstructed permutation dataset. Under the continuous no-ties model, an event at $T = t$ has landmark pair $(0, 1)$ and a censoring at $T = t$ has pair $(1, 0)$; these boundary observations are handled explicitly. For dataset index $j = 0, \ldots, B$, we use the raw upper-tailed maximum

$$\Lambda_s^{(j)} = \max_k \widehat{\Delta I}^{\,s}_{t_{jk}}(D^{(j)}),$$

where $j = 0$ identifies the original dataset. Multiplying each stratum's widths by its fixed positive weight n_s/n leaves its ranks unchanged. The stratum ranks, Fisher aggregation, and final global rank are calculated symmetrically across all $B + 1$ datasets using Algorithm 2.

Oracle conditional-completion control. To distinguish the randomization step from learned imputation, we also apply the same statistic and permutation procedure using missing times sampled from the known conditional null distribution. This control uses truncated Weibull margins or the conditional event-only frailty mixture and retains every observed time. It uses generating distributions, not the realized hidden event and censoring times, to sample the missing values. The oracle is a diagnostic reference; its additional information is not available to the learned detector.

H.3 RESULTS

For each setting, Table 7 reports the number of false rejections under the strict rule $p_{\text{global}} < 0.05$, the empirical false-positive rate (FPR), and a two-sided 95% Clopper–Pearson interval. We additionally report

$$p_{\text{Fisher}} = \Pr\left\{ \chi^2_{2R} \geq -2 \sum_{r=1}^{R} \log p_r \right\}$$

as a nominal combined-evidence summary across the R independent datasets. This summary is distinct from the FPR and does not imply that each individual replication is a non-rejection.

Table 7: Null-calibration results for the proposed MI-uncertainty test with the raw upper-tailed maximum. Each row uses $R = 10$ independently generated datasets of size $n = 1000$. FPR uses the strict rule $p_{\text{global}} < 0.05$; intervals are two-sided 95% Clopper–Pearson intervals. The oracle column replaces only conditional completion, keeping the statistic and testing procedure unchanged.

Null setting	False rejects/R	FPR	95% interval	p_{Fisher}	Oracle rejects/R
Clayton, $\theta = 0$	0/10	0.00	$[0.000, 0.308]$	0.8149	0/10
Frank, $\theta = 0$	1/10	0.10	$[0.003, 0.445]$	0.4792	1/10
Gaussian, $\rho = 0$	1/10	0.10	$[0.003, 0.445]$	0.4016	0/10
Gumbel, $\theta = 1$	0/10	0.00	$[0.000, 0.308]$	0.5296	0/10
Frailty, $(\alpha_E, \alpha_C) = (0, 0)$	1/10	0.10	$[0.003, 0.445]$	0.1791	0/10
Frailty, $(\alpha_E, \alpha_C) = (1, 0)$	1/10	0.10	$[0.003, 0.445]$	0.2014	0/10

The proposed test produces zero or one false rejection in each batch of ten datasets, for a total of $4/60 \simeq 0.067$. The oracle control produces $1/60$ false rejections. These totals are descriptive summaries of the six null batches, not additional independent experiments or a common-rate estimate across different null mechanisms. All per-setting confidence intervals include the nominal level 0.05, although they are wide because $R = 10$. The study provides preliminary empirical support for low false-positive behavior under the specified null designs and training configuration. It complements the conditional calibration theory without claiming uniform Type-I-error control for every generating law or training choice.