PROPER SCORING RULES FOR RIGHT-CENSORED SURVIVAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Proper scoring rules provide a rigorous theoretical basis for the training and evaluation of probabilistic forecasts. In survival analysis, such forecasts describe the distribution of the time until an event occurs. However, this event time is often only partially observed because follow-up may end before the event occurs, resulting in right censoring. We propose a framework for proper scoring of right-censored survival outcomes based on a simple idea: first, map the predictive distribution through the censoring mechanism, then apply the underlying proper score on the induced observed-data law. This yields localized scores for fixed censoring times and marginalized scores when the censoring time is random or only partially observed. The resulting construction recovers familiar right-censored likelihood and IPCW-type criteria within a coherent framework, while also yielding right-censored versions of the CRPS, pinball loss, Brier score, and energy score. We show that the marginalized construction is proper under conditional independent censoring and, for strictly proper base scores, identifies the latent joint CDF on the identifiable region. The same principle also leads to censored engression, a sample-based learning objective for multivariate right-censored survival modeling. In experiments, our scores correctly rank the oracle forecast across several censoring regimes, whereas forecast-dependent plug-in weighted scores can exhibit ranking reversals. Censored engression likewise substantially improves over naive training on censored outcomes.

1 INTRODUCTION

Predicting time-to-event outcomes is central in many applications, including healthcare, reliability, customer churn, and risk modeling. In these settings, the goal is often not merely to estimate the mean survival time or a hazard, but to forecast the full conditional event-time distribution F_X given covariates X. Proper scoring rules (Gneiting & Raftery, 2007) provide principled criteria for training and evaluating probabilistic forecasts, and they ensure that the true predictive distribution is optimal in expectation.

The main difficulty in survival prediction is that event times are only partially observed. Under right censoring, one observes $Y = \min(T, C)$ and $\Delta = \mathbb{1}\{T \leq C\}$, rather than the latent event time T itself, where C denotes the censoring time. As a result, standard proper scoring rules for uncensored outcomes cannot be applied directly. This raises a basic question: how should one define a proper scoring rule when the event time is coarsened by right censoring?

Our answer is to score forecasts after applying the same coarsening operation that generated the data. Given a predictive distribution F for T and a censoring time c, we evaluate the induced distribution of $(\min(T, c), \mathbb{1}\{T \leq c\})$ under $T \sim F$, rather than evaluating F against the unobserved T. Thus, the score is determined by the observation mechanism: the censoring map is fixed independently of the forecast being evaluated. For fixed c, this gives a localized proper score in the sense of de Punder et al. (2026). For random censoring, the realized censoring time may be only partially observed; for example, if an event is observed, one only knows that $C \geq Y$. We therefore average the localized score over the conditional law of C given (Y, Δ, X) under the censoring mechanism. The resulting marginalized score remains forecast-independent once the censoring law is fixed, or estimated separately from the forecast being evaluated, in contrast to plug-in constructions (Yanagisawa, 2023) where censoring weights or tail probabilities are recomputed from the forecast itself.

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.

Beyond evaluation, the same principle naturally leads to a training objective for flexible survival models. We extend engression (Shen & Meinshausen, 2025), an implicit distributional regression method based on the energy score, to right-censored multivariate survival data. The resulting *censored engression* method trains a noise-conditioned generator by comparing censored generated samples with observed censored outcomes, using the same localized or marginalized right-censored energy score. This makes it possible to learn flexible joint event-time distributions under censoring without requiring a tractable likelihood.

Contributions. First, building on localization by censoring the predictive distribution prior to scoring (de Punder et al., 2026), we present a unified construction of proper scoring rules for right-censored outcomes, including settings in which the stochastic censoring time is only partially observed. Second, for random or partially observed censoring, we introduce forecast-independent marginalization over compatible censoring times, preserving propriety and, for strictly proper base scores, identifying the latent joint CDF on the identifiable region. Third, we derive concrete right-censored versions of the logarithmic score, CRPS, pinball loss, Brier score, and energy score, clarifying their relationship to existing survival criteria (see Table 1). Fourth, we use the censored energy score as a training objective to define *censored engression* for flexible multivariate right-censored survival modeling.

Finally, our experiments separate three empirical questions. In controlled settings with known data-generating processes (DGPs), we show that our forecast-independent scores correctly rank the oracle forecast across censoring regimes, whereas forecast-dependent plug-in weighted scores can exhibit ranking reversals. In synthetic multivariate learning experiments, we demonstrate that censored engression outperforms naive training and is competitive with likelihood-based Weibull, copula, and low-dimensional grid benchmarks trained using censored log scores. In an intensive care unit (ICU) acute kidney injury (AKI) case study, we show that censored engression can flexibly model joint time-to-event distributions for clinically relevant creatinine- and urine-output-based AKI endpoints.

2 Related work

Our framework is closely related to localization and coarsening. de Punder et al. (2026) show that a proper scoring rule can be localized by applying it to a censored predictive distribution; our fixed-censoring construction is an application of this principle to right-censored survival outcomes. More broadly, Gill et al. (1997) develop the coarsening-at-random framework for partially observed data. Our main extension begins when the censoring time is stochastic or only partially observed: we average the localized score over censoring times compatible with the observation using the censoring law, while keeping this adjustment fixed independently of the candidate forecast.

For right-censored survival evaluation, Graf et al. (1999) introduced the IPCW Brier score and its integrated version, and Gerds & Schumacher (2006) established consistency under appropriate censoring models, including covariate-conditional censoring under conditional independence. Kvamme & Borgan (2023) further study difficulties of standard IPCW evaluation under administrative censoring when censoring times are predictable from covariates. Avati et al. (2020) proposed the Survival-CRPS for censored outcomes, while Rindt et al. (2022) later showed that this score and standard integrated Brier implementations need not be proper and established propriety of the usual right-censored log-likelihood; our construction recovers the latter and, with the appropriate conditional censoring law, the IPCW Brier score as special cases. Alberge et al. (2025) propose a strictly proper censoring-adjusted separable score for survival and competing risks, with a focus on efficient stochastic optimization. Their competing-risks formulation concerns mutually exclusive event types, whereas our multivariate setting allows event times to co-occur.

Censored quantile estimation has also been studied extensively. Portnoy (2003) and Pearce et al. (2022) develop censored quantile procedures based on reweighting censored observations using information about the event-time distribution. Among score-specific approaches, Yanagisawa (2023) formulates related censored pinball, Brier, and ranked probability scores and proves propriety when the required weights use the correct event-time law. Our marginalized construction instead integrates the localized score over the censoring law, which is held fixed across candidate forecasts. Thus, while fixed-censoring localization and several resulting scores are closely related to existing work,

the forecast-independent marginalization over random or partially observed censoring is our main theoretical contribution.

Modern neural survival models span discrete-time competing-risk approaches such as DeepHit (Lee et al., 2018), which combines a right-censoring-adjusted likelihood for the first event and its cause with cause-specific ranking losses, and continuous-time likelihood models such as Deep Survival Machines (Nagpal et al., 2021), conditional normalising flows (Ausset et al., 2021), and survival mixture-density networks (Han et al., 2022), which obtain flexibility through neural parametric mixtures or tractable density transformations. Adversarial approaches instead learn event-time or survival-data distributions through generative objectives, as in adversarial time-to-event modelling (Chapfuwa et al., 2018) and SurvivalGAN (Norcliffe et al., 2023), although the latter primarily targets synthetic survival-data generation rather than conditional survival forecasting. In contrast, censored engression learns an implicit distribution over potentially co-occurring event times by applying the censoring mechanism to generated samples and optimising a localized or marginalized proper energy score, avoiding a tractable likelihood while identifying the latent joint CDF only on the region supported by the censoring mechanism.

3 Proper scoring under censoring

The development below separates an inherited ingredient from our main extension. Section 3.1 specializes the localization principle of de Punder et al. (2026) to right-censored survival outcomes with a fixed censoring configuration. Our main theoretical extension is the marginalized construction for random or partially observed censoring times, introduced subsequently, for which Proposition 1 gives a single propriety argument applicable to any suitable localized proper score. Deriving tractable formulas for particular scores remains score-specific.

We propose a general framework for proper scoring rules for right-censored survival outcomes that recovers, up to a forecast-independent positive factor, previously proposed proper scoring rules, such as the censored logarithmic score (Rindt et al., 2022) and the IPCW Brier score (Graf et al., 1999; Gerds & Schumacher, 2006). Our starting point is a strictly proper scoring rule on the latent outcome space. Following de Punder et al. (2026), we view censoring as a transformation of the predictive distribution itself: rather than scoring the latent law directly, we first map it to the censored outcome space and then apply the corresponding proper score to the induced observed-data law. This preserves the original score divergence after censoring and yields a score that is locally proper for the latent law on the part of the outcome space that remains identifiable under the censoring mechanism.

Let $X \in \mathcal{X} \subseteq \mathbb{R}^d$ denote covariates, and let $T \in \mathcal{T} \subseteq \mathbb{R}_+^k$ denote the latent event-time outcome, with $k = 1$ in the univariate case and $k \geq 2$ in the multivariate setting. For each covariate value x, we seek a predictive distribution $F = \mathcal{L}(T \mid X = x)$. Under right censoring, however, we observe only $Y = \min(T, C)$ and $\Delta = \mathbb{1}\{T \leq C\}$.

Equip $\mathcal{T}$ with a σ-algebra $\mathscr{T}$, let $\mathcal{P}$ be a class of probability measures on $(\mathcal{T}, \mathscr{T})$, and let $S : \mathcal{P} \times \mathcal{T} \to \bar{\mathbb{R}}$ be a scoring rule with divergence

$$D_S(P\|F) = \mathbb{E}_P[S(F,T)] - \mathbb{E}_P[S(P,T)].$$

We call S strictly proper if $D_S(P\|F) \geq 0$ for all $P, F \in \mathcal{P}$, with equality only when $P = F$. This is the full-data scoring rule that would be used if the latent event time T were observed. Although it cannot generally be evaluated under censoring, it serves as the reference objective from which we construct the censored scores below.

3.1 Localized proper scoring rules under fixed censoring

We begin with a fixed censoring configuration. Let $A \subseteq \mathcal{T}$ denote the part of the outcome space that remains identifiable under this censoring mechanism, and let $*$ denote an abstract censored outcome. The corresponding censored outcome space is $\mathcal{T}_A^\flat = A \cup \{*\}$. For a latent predictive law $F \in \mathcal{P}$, define its censored version by

$$F_A^\flat = F_A + \bar{F}_A\, \delta_*, \qquad F_A(B) := F(B \cap A), \qquad \bar{F}_A := F(A^c).$$

Thus, the law is left unchanged on the identifiable region A, while all mass outside A is collapsed into the censored outcome $*$.

By a slight abuse of notation, we use S also for the corresponding proper scoring rule on the censored outcome space $\mathcal{T}_A^\flat$. The localized score is then

$$S_A^\flat(F,t) := S(F_A^\flat, t_A^\flat), \qquad \text{where} \quad t_A^\flat = \begin{cases} t, & t \in A, \\ *, & t \in A^c. \end{cases}$$

Intuitively, following the localization perspective of de Punder et al. (2026), the forecast is first censored in the same way as the observation, and the original proper score is then evaluated on the resulting censored outcome space. By construction, the induced divergence is the original divergence applied to the censored laws, $D_{S_A^\flat}(P\|F) = D_S(P_A^\flat\|F_A^\flat)$, so equality of population score implies equality of the induced censored distributions, not necessarily of the full latent laws. Thus, strictness applies to the induced censored law; recovery of the latent law is limited to what the censoring mechanism identifies.

For fixed right censoring at time c, the identifiable region depends on the boundary convention. With the convention $\Delta = \mathbb{1}\{T \le c\}$, the identifiable region is $A_c = [0, c]$, and the censored outcome map is

$$\psi_c^\flat(t) = \begin{cases} t, & t \le c, \\ *, & t > c. \end{cases}$$

The corresponding censored law is $F_c^\flat = F(\cdot \cap [0,c]) + F((c,\infty))\delta_*$. Thus, an event at the boundary c is distinguished from a censored observation beyond c. We write $S_c^\flat(F; Y, \Delta)$ for the localized score induced by this fixed censoring time.

When $T \mid X$ is continuous, boundary events have probability zero, so the convention at $T = c$ does not affect expected scores. For distance-based scores such as the CRPS and energy score, however, the abstract censored state $*$ cannot be inserted directly into Euclidean distances. By a slight abuse of notation, we keep the notation $\psi_c^\flat$ for this numerical encoding of the same coarsened observation:

$$\psi_c^\flat(t) = \min(t, c), \qquad F_c^\flat = \mathcal{L}(\psi_c^\flat(T)).$$

This representation should be read as an encoding of the coarsened observation: the value c represents the censored region (c, ∞), not an event known to occur at c. When the event indicator is needed, we keep the convention $\Delta = \mathbb{1}\{T \le c\}$, so that $(c, 0)$ denotes censoring beyond c whereas $(c, 1)$ denotes an event observed at the boundary.

The single-atom representation above describes scalar censoring. In the multivariate setting, componentwise censoring retains which coordinates are censored, and we therefore work directly with the componentwise censoring map,

$$\psi_c^\flat(t) = (t_1 \wedge c_1, \ldots, t_k \wedge c_k), \qquad F_c^\flat = \mathcal{L}(\psi_c^\flat(T)),$$

where $c = (c_1, \ldots, c_k)$. A shared scalar censoring time corresponds to the special case $c_1 = \cdots = c_k = c$.

3.2 Marginalized proper scoring rules

We now turn from a fixed censoring threshold to random right censoring. Suppose that $Y = \min(T, C)$, $\Delta = \mathbb{1}\{T \le C\}$, and let $H(t \mid x) = \mathbb{P}(C \le t \mid X = x)$ denote the conditional censoring distribution. To relate scoring of the observed data (Y, Δ) to the latent conditional event-time distribution $T \mid X$, we assume conditional independent censoring.

Assumption 1 (Conditional independent censoring)**.** $T \perp C \mid X$.

Under Assumption 1, the censoring mechanism carries no information about the latent event time beyond the observed covariates. Importantly, this assumption is distinct from whether the realized censoring time is observed. In the usual right-censoring setting, C may itself be only partially observed: if $\Delta = 1$, for example, the event is observed at $Y = T$ and one only learns that $C \ge Y$.

Let $\Pi(dc \mid Y, \Delta, X) := \mathcal{L}(C \in dc \mid Y, \Delta, X)$ denote the conditional law of the censoring time compatible with the observed data, induced by $H(\cdot \mid X)$ and the right-censoring mechanism. We define the observed-data score by averaging the fixed-c localized score over this conditional law:

$$\bar{S}(F; Y, \Delta, X) := \int S_c^\flat(F; Y, \Delta)\, \Pi(dc \mid Y, \Delta, X).$$

Thus, the score first conditions on each censoring configuration compatible with the observation and then averages over those configurations according to the censoring law, which is fixed independently of the forecast F.

In the univariate case, this reduces to

$$\bar{S}(F; Y, \Delta, X) = \begin{cases} \displaystyle\int_{[Y,\infty)} S_c^\flat(F; Y, 1)\, H(dc \mid C \geq Y, X), & \Delta = 1, \\ S_Y^\flat(F; Y, 0), & \Delta = 0. \end{cases}$$

When $\Delta = 0$, the censoring time is observed exactly as $C = Y$; when $\Delta = 1$, it is only known to satisfy $C \geq Y$. Similarly, with a shared censoring time in the multivariate case, C is observed whenever at least one coordinate is censored. If all coordinates are observed, one only knows that $C \geq \max_j Y_j$, and Π is the law of C conditional on this event.

The censoring mechanism also determines which parts of the latent distribution are identifiable. For fixed x, let

$$\bar{\mathcal{S}}_x(F) := \mathbb{E}_P[\bar{S}(F; Y, \Delta, X) \mid X = x]$$

denote the conditional expected score. For a shared scalar censoring time, define

$$\mathcal{I}_x^{\text{shared}} = \left\{ t \in \mathbb{R}_+^k : \mathbb{P}\left(C > \max_{j=1,\ldots,k} t_j \mid X = x \right) > 0 \right\},$$

whereas for componentwise censoring times $C = (C_1, \ldots, C_k)$, define

$$\mathcal{I}_x^{\text{comp}} = \left\{ t \in \mathbb{R}_+^k : \mathbb{P}(C_j > t_j \text{ for all } j = 1, \ldots, k \mid X = x) > 0 \right\}.$$

We write $\mathcal{I}_x$ for the relevant identifiable region under the censoring mechanism considered. In the univariate case, both definitions reduce to $\mathcal{I}_x = \{t \geq 0 : \mathbb{P}(C > t \mid X = x) > 0\}$.

The following proposition shows that averaging over the compatible censoring configurations preserves propriety, while strict propriety can hold only on the region that is identifiable under the censoring mechanism.

Proposition 1 (Marginalized propriety under right censoring)**.** *Assume Assumption 1, and let $H(\cdot \mid X)$ denote the true conditional distribution of the censoring time. Suppose that the conditional law of C given (Y, Δ, X) used in the marginalized score is the one induced by $H(\cdot \mid X)$ and the right-censoring mechanism. If, for $H(\cdot \mid X = x)$-almost every censoring time c, the localized score $S_c^\flat$ is proper for the induced censored law, then, for every x, $\bar{\mathcal{S}}_x(F) \geq \bar{\mathcal{S}}_x(P)$ for all forecasts F.*

If, in addition, the localized scores are strictly proper, then

$$\bar{\mathcal{S}}_x(F) = \bar{\mathcal{S}}_x(P) \quad \Longleftrightarrow \quad F_{c,x}^\flat = P_{c,x}^\flat \quad \text{for } H(\cdot \mid x)\text{-almost every } c.$$

Moreover, equality of the marginalized scores implies

$$F_x(t) = P_x(t) \qquad \text{for every } t \in \mathcal{I}_x,$$

where F_x and P_x denote the joint conditional CDFs.

In particular, if $\mathcal{I}_x = \mathbb{R}_+^k$, then $F_x = P_x$ as probability laws.

Outside $\mathcal{I}_x$, two latent distributions may differ while inducing the same observed censored law, so no scoring rule based only on (Y, Δ, X) can be strictly proper for such non-identifiable differences. This matters whenever a downstream quantity depends on the distribution beyond supported follow-up: for example, under administrative follow-up to two years, a five-year event probability necessarily relies on extrapolation. The same applies to upper quantiles, mean survival time, and other tail-dependent functionals.

The conclusion of Proposition 1 also relies on Assumption 1. If conditional independent censoring fails, the marginalized score remains a well-defined observed-data criterion under the specified censoring mechanism, but its population minimizer need not equal $\mathcal{L}(T \mid X)$. For example, an unobserved factor may affect both T and C. Without additional assumptions or a joint model for (T, C), the resulting distortion is generally not identifiable from (Y, Δ, X) alone.

Table 1: Examples of uncensored, fixed-censoring localized, and marginalized scores under right censoring. All scores are written as losses, so lower values are better. We write $G(t \mid x) = \mathbb{P}(C > t \mid X = x)$ and $G(t^- \mid x) = \mathbb{P}(C \geq t \mid X = x)$.

Type	Score
Logarithmic score (log)	
Uncensored	$S_{\log}(F; t, x) = -\log f(t \mid x)$
Fixed local	$S^{\flat}_{\log,c}(F; y, \delta, x) = -\delta \log f(y \mid x) - (1-\delta)\log(1 - F(c \mid x))$
Marginalized	$\bar{S}_{\log}(F; y, \delta, x) = -\delta \log f(y \mid x) - (1-\delta)\log(1 - F(y \mid x))$
Continuous ranked probability score (CRPS)	
Uncensored	$S_{\mathrm{CRPS}}(F; t, x) = \int_0^t F(s \mid x)^2\, ds + \int_t^\infty (1 - F(s \mid x))^2\, ds$
Fixed local	$S^{\flat}_{\mathrm{CRPS},c}(F; y, \delta, x) = \int_0^y F(s \mid x)^2\, ds + \delta \int_y^c (1 - F(s \mid x))^2\, ds$
Marginalized	$\bar{S}_{\mathrm{CRPS}}(F; y, \delta, x) = \int_0^y F(s \mid x)^2\, ds + \delta \int_y^\infty \frac{G(s \mid x)}{G(y^- \mid x)} (1 - F(s \mid x))^2\, ds$
Brier score at horizon τ ($\mathbf{BS}_\tau$)	
Uncensored	$S_{\mathrm{BS},\tau}(F; t, x) = \big(F(\tau \mid x) - \mathbb{1}\{t \leq \tau\}\big)^2$
Fixed local	$S^{\flat}_{\mathrm{BS},\tau,c}(F; t, x) = \big(F^{\flat}_c(\tau \mid x) - \mathbb{1}\{\min(t,c) \leq \tau\}\big)^2$
Marginalized	$\bar{S}_{\mathrm{BS},\tau}(F; Y, \Delta, X) = \mathbb{1}\{Y > \tau\} F(\tau \mid X)^2 + \Delta \mathbb{1}\{Y \leq \tau\} \frac{G(\tau \mid X)}{G(Y^- \mid X)} (1 - F(\tau \mid X))^2$
Pinball / quantile score at level α ($\mathbf{Q}_\alpha$)	
Uncensored	$S_{\mathrm{Q},\alpha}(F; t, x) = \big(\alpha - \mathbb{1}\{t < q_\alpha(F \mid x)\}\big)\big(t - q_\alpha(F \mid x)\big)$
Fixed local	$S^{\flat}_{\mathrm{Q},\alpha,c}(F; t, x) = \big(\alpha - \mathbb{1}\{\min(t,c) < q_\alpha(F^{\flat}_c \mid x)\}\big)\big(\min(t,c) - q_\alpha(F^{\flat}_c \mid x)\big)$
Marginalized	$\bar{S}_{\mathrm{Q},\alpha}(F; Y, \Delta, X) = \alpha(Y - q_\alpha(F \mid X))_+ + \Delta \frac{1-\alpha}{G(Y^- \mid X)} \mathbb{1}\{q_\alpha(F \mid X) > Y\} \int_Y^{q_\alpha(F \mid X)} G(t \mid X)\, dt$
Energy score (ES)	
Uncensored	$S_{\mathrm{ES},\beta}(F; t, x) = \mathbb{E}[\|Z - t\|^\beta \mid X = x] - \frac{1}{2}\mathbb{E}[\|Z - Z'\|^\beta \mid X = x]$
Fixed local	$S^{\flat}_{\mathrm{ES},\beta,c}(F; y, x) = \mathbb{E}[\|\psi_c(Z) - y\|^\beta \mid X = x] - \frac{1}{2}\mathbb{E}[\|\psi_c(Z) - \psi_c(Z')\|^\beta \mid X = x]$
Marginalized	$\bar{S}_{\mathrm{ES},\beta}(F; Y, \Delta, X) = \mathbb{E}\Big[S^{\flat}_{\mathrm{ES},\beta,C}(F; Y, X) \mid Y, \Delta, X\Big]$

In the univariate continuous case, taking the logarithmic score as the base score recovers the usual right-censored negative log-likelihood; see Section B.1 and Table 1. In this case, marginalization simplifies because the fixed-c localized contribution is either observed exactly or is constant over the censoring times compatible with the observation.

Table 1 summarizes several scores obtained from the localized and marginalized constructions. The formulas use the continuous no-boundary-ambiguity convention, under which the abstract censored outcome can be represented by the boundary value c through $\psi^{\flat}_c(t) = \min(t, c)$. Detailed derivations are given in Appendix B.

Remark 1 (Fixed, administrative, and observed censoring times)**.** *If $C = c$ is fixed by design, or $C = c(X)$ is deterministic given the observed covariates, the localized score can be evaluated directly and no marginalization is required; conditional independent censoring then holds automatically.*

If C is stochastic but observed for every individual, marginalization is again unnecessary: the localized score can be evaluated at the observed C. Observing C, however, does not imply Assumption 1; latent factors may affect both T and C, so that $T \not\perp C \mid X$. Under conditional independent censoring, the expected localized score is proper on the identifiable region. This setting is relevant, for example, when discharge or death times are recorded in the ICU application of Section 5.3.

Estimating the censoring law. The construction assumes that the censoring law $H(\cdot \mid X)$ is known or estimated independently of the forecast being evaluated. When it is not known from the study design, it can be estimated on a separate calibration sample or by cross-fitting the evaluation data. In the latter case, the censoring model is fitted on $K - 1$ folds and used to score observations in the held-out fold. The same fitted censoring models and folds are used for all candidate forecasts,

which remain fixed throughout; in particular, the censoring adjustment is never recomputed from the forecast being evaluated.

4 CENSORED ENGRESSION FOR MULTIVARIATE SURVIVAL

Thus far, we have used censoring-aware proper scoring rules to *evaluate* predictive distributions under right censoring. The same scores can also serve as learning objectives. We develop this connection for the energy score by combining two ideas: the implicit distributional representation of engression (Shen & Meinshausen, 2025) and the localization-by-censoring principle of Section 3.1. The resulting *censored engression* method learns flexible, possibly multivariate, event-time distributions by applying the censoring map to generated samples before evaluating the energy score, with marginalization over compatible censoring configurations when necessary.

Background on engression. Engression (Shen & Meinshausen, 2025) represents the conditional law through a generator

$$Y_g(x) = g(x, \varepsilon), \qquad \varepsilon \sim P_\varepsilon, \tag{1}$$

where P_ε is a simple reference distribution. The generator induces the conditional predictive law $P_g(\cdot \mid x) := \mathcal{L}(g(x, \varepsilon))$, so unlike classical regression, engression directly models the full conditional distribution via samples. For two independent noise variables $\varepsilon, \varepsilon'$, the energy score population objective is

$$\mathcal{R}_{\mathrm{ES}}(g) = \mathbb{E}\left[\|Y - g(X, \varepsilon)\| - \frac{1}{2}\|g(X, \varepsilon) - g(X, \varepsilon')\|\right],$$

where $(X, Y) \sim P$. Under correct specification, minimizing this objective recovers the true conditional law; see Proposition 1 of Shen & Meinshausen (2025).

4.1 CENSORED ENGRESSION

We now extend engression to multivariate right-censored survival outcomes. Let $T = (T_1, \ldots, T_k) \in \mathbb{R}_+^k$, $k \geq 1$, denote the latent event-time vector, and let the generator in equation 1 induce the latent predictive law $F_{g,x} := \mathcal{L}(g(x, \varepsilon))$.

Under right censoring, the event times are not fully observed. For a fixed censoring time c, let $\psi_c^\flat(t) = \big(\min(t_1, c), \ldots, \min(t_k, c)\big)$ denote the shared censoring map, and let $Y = \psi_c^\flat(T)$ denote the corresponding censored outcome. Following the localization-by-censoring principle, we apply the energy score after censoring, that is, to the induced censored law $\mathcal{L}(\psi_c^\flat(g(x, \varepsilon)))$ rather than to the latent law $F_{g,x}$ itself. The resulting fixed-c censored energy loss is

$$\mathcal{R}_{\text{local-cens-ES}}(g; Y, x, c) = \mathbb{E}_\varepsilon\left[\|\psi_c(g(x, \varepsilon)) - Y\|\right] - \frac{1}{2}\mathbb{E}_{\varepsilon, \varepsilon'}\left[\|\psi_c(g(x, \varepsilon)) - \psi_c(g(x, \varepsilon'))\|\right], \tag{2}$$

where $\varepsilon, \varepsilon' \overset{\text{iid}}{\sim} P_\varepsilon$.

When the censoring time is not fully observed, we average the fixed-c loss over the compatible conditional law of C given the observed data and then over the observed-data distribution:

$$\mathcal{R}_{\text{cens-ES}}(g) := \mathbb{E}[\mathbb{E}[\mathcal{R}_{\text{local-cens-ES}}(g; Y, X, C) \mid Y, \Delta, X]]. \tag{3}$$

The inner conditional expectation is exactly the observation-level marginalized right-censored energy score introduced above, while $\mathcal{R}_{\text{cens-ES}}(g)$ is its population expectation. When there is no censoring, ψ_c is the identity map and the loss reduces to the ordinary engression energy score. With censoring, the model is instead trained to match the law of the censored outcome, and therefore recovers the latent event-time law only to the extent that it remains identifiable under the censoring mechanism. We refer to this procedure as *censored engression*. We also present this approach with a population optimality guarantee; see Proposition 2.

5 EXPERIMENTS

5.1 PROPER SCORING RULES FOR EVALUATING FORECASTS

We use controlled simulations with known data-generating laws to test whether the proposed scores preserve the oracle ranking under censoring. We compare our forecast-independent scores with scores of Yanagisawa (2023) using either oracle weights or plug-in weights recomputed from each candidate forecast; full simulation details are given in Appendix C.2.

Table 2: Rank of the oracle forecast. Rank 1 means best mean score. Regimes A–C compare F_0 against $F_1, \ldots, F_4$; regime D compares $F_0^{(D)}$ against four exploit forecasts $F_5(\epsilon)$.

Scoring rule	A	B	C	D
CRPS, latent	1	1	1	1
Pinball, latent	1	1	1	1
Brier, latent	1	1	1	1
NLL, latent	1	1	1	1
NLL (Yanagisawa, 2023), plug-in	2	1	2	2
Brier MC (Yanagisawa, 2023), plug-in	1	1	1	5
Brier (Yanagisawa, 2023), plug-in	1	1	2	5
Pinball (Yanagisawa, 2023), plug-in	1	1	1	1
RPS (Yanagisawa, 2023), plug-in	2	1	2	5
Censored NLL, ours	1	1	1	1
CRPS, ours	1	1	1	1
Pinball, ours	1	1	1	1
Brier, ours	1	1	1	1

Results and discussion. Table 2 shows that, across all four censoring regimes, the latent gold-standard scores and all of our forecast-independent scores rank the oracle first. The oracle-weight scores of Yanagisawa (2023) likewise rank the oracle first, whereas forecast-dependent plug-in versions can reverse the ranking: the plug-in NLL ranks the oracle second in regimes A, C, and D, the plug-in RPS ranks it second in A and C and last in D, and plug-in Brier variants also fail in C or D. Corresponding raw score values are reported in Appendix C.3. Sensitivity analyses in Appendix C.4 show that the ranking is largely stable when the censoring law is estimated rather than known, with only mild degradation under misspecification.

5.2 Synthetic multivariate censored-engression experiment

We next study censored engression on multivariate event times $T = (T_1, \ldots, T_k)$, $k \in \{2, 3, 5, 10\}$, under a shared censoring time. The main DGP is a covariate-dependent mixture log-normal distribution, yielding multimodal conditional laws outside the Weibull-copula families used by the likelihood baselines. We compare censored engression with naive observed engression, neural Weibull and copula models trained by censored likelihood, and low-dimensional Gumbel-copula and grid baselines; full details are in Appendix C.5. Common competing-risk models such as DeepHit and Deep Survival Machines do not directly model our target because event times may co-occur; fitting them separately for each endpoint would discard their joint dependence.

Results. Censored energy score is our primary observed-data criterion, with latent energy score serving as a synthetic diagnostic. Censored engression obtains the lowest censored energy score among feasible methods across the reported censoring regimes and dimensions. Under uniform censoring, for example, it improves over naive observed engression from 0.5524 to 0.4904 at $k = 2$ and from 1.6402 to 1.3382 at $k = 10$. At $k = 2$, the strongest likelihood baseline, MLP Clayton–Weibull, obtains 0.5152. Likelihood-based models can nevertheless achieve lower latent energy score, reflecting behavior beyond the region identified by censored observations. Full results and the unimodal sanity check are reported in Appendix C.6.

5.3 Illustrative clinical use case: early AKI forecasting in the ICU

We use early acute kidney injury (AKI) forecasting in the intensive care unit (ICU) as a realistic illustration of censored engression on multivariate time-to-event data. At each prediction time, the goal is to forecast the joint distribution of the remaining time to two clinically meaningful endpoints: creatinine-defined AKI and urine-output-defined AKI. The predictors are routinely available variables, including demographics, baseline renal function, elapsed ICU time, and summary features of recent creatinine and urine-output measurements over the preceding 72 hours. The analysis uses de-identified ICU data from MIMIC-IV (Johnson et al., 2023); additional cohort, endpoint, and evaluation details are given in Appendix C.7.

The observed target is right-censored by ICU discharge or death. This exit time is observed in the database, so the localized version of our scoring framework can be used directly: for each patient and prediction time, generated event times and observed outcomes are both mapped through the same censoring time before the score is evaluated. This removes the need to marginalize over unknown censoring times, but does not imply conditional independent censoring. We assume that the

Table 3: Selected results for censored engression. Lower scores are better. Synthetic rows report both observed censored energy score and latent energy score under uniform censoring for the mixture log-normal DGP. DGP is the true data-generating law and latent engression is an infeasible reference trained on uncensored event times. AKI rows report localized energy scores; no latent oracle is available in the real-data use case. Bold indicates the best feasible result among the displayed methods.

Setting	Score	DGP	Latent engression	Naive obs. engression	Censored engression
Synthetic, $k = 2$, uniform	Censored ES	0.4840	0.4907	0.5524	**0.4904**
	Latent ES	1.2468	1.2601	1.4440	**1.2887**
Synthetic, $k = 10$, uniform	Censored ES	1.3131	1.6082	1.6402	**1.3382**
	Latent ES	9.4613	10.0111	16.5122	**14.2576**
AKI, joint	Loc. ES	–	–	43.442	**30.933**
AKI, creatinine	Loc. ES	–	–	23.591	**15.439**
AKI, urine-output	Loc. ES	–	–	32.981	**23.769**

latent AKI times and ICU exit time are conditionally independent given the available predictors; if residual dependence remains, for example through unmeasured clinical severity, the results should be interpreted as predictive performance under the observed exit mechanism rather than as identification of latent post-exit AKI incidence.

We report localized energy scores for the joint outcome and each marginal endpoint, together with localized Brier scores for "either AKI" at clinically relevant horizons. We focus on proper distributional scores rather than the C-index, which evaluates discrimination after reducing the multivariate prediction to an endpoint or scalar risk score. Table 3 compares censored engression with a naive observed-engression baseline using the same generator architecture.

Censored engression improves all localized energy scores in Table 3: the joint score decreases from 43.442 to 30.933, the creatinine-AKI score from 23.591 to 15.439, and the urine-output-AKI score from 32.981 to 23.769. Localized Brier scores for "either AKI" likewise decrease from 0.023 to 0.022 at 6 hours, 0.045 to 0.040 at 12 hours, and 0.081 to 0.071 at 24 hours.

Appendix C.7 compares censored engression with standard survival baselines. It obtains the lowest localized Brier scores across the reported horizons (Appendix Table 14) and the lowest joint and marginal localized energy scores among methods that generate full event-time distributions (Appendix Table 15).

6 Conclusion

This work argues that censored probabilistic evaluation should be built on forecast-independent censoring adjustment. IPCW scores target the latent event-time distribution only when the weights are determined by the censoring mechanism rather than by the forecast being evaluated. This preserves the oracle ranking across a range of censoring regimes, while forecast-dependent plug-in scores change the evaluation target and can induce ranking reversals.

The same methodological principle yields a training objective for censored generative survival modeling. Censored engression uses the censored energy score to extend engression to right-censored multivariate outcomes, improving over naive training on censored observations.

An important extension is to more general coarsening mechanisms, including interval censoring. Recent work by Yanagisawa & Akiyama (2026) develops a strictly proper interval-censored log score under additional censoring assumptions; extending the localization-and-marginalization perspective developed here to broader classes of proper scores under interval and set censoring remains an interesting direction.

Limitations. Our guarantees rely on conditional independent censoring and a correctly specified censoring law. If $T \not\perp C \mid X$, the score need not target the marginal latent event-time law. Strict propriety is also limited to the identifiable region, so long-horizon risks and other tail-dependent quantities may require extrapolation. In practice, the censoring law should be estimated separately and then held fixed across candidate forecasts; misspecification can still affect ranking.

AI USE STATEMENT

In this work, we used generative AI tools for language-related assistance, including grammar correction, wording refinement, and text styling. We also used generative AI tools to assist with code generation, debugging, and general programming-related questions, as well as to support parts of the literature research process, such as identifying potentially relevant papers and research directions. We have not used generative AI tools for any other tasks requiring disclosure, and other disclosure categories are not applicable to this work.

We reviewed and verified all AI-assisted work. AI-assisted text was manually reviewed and edited by the authors. AI-generated or AI-assisted code was inspected, tested, and verified for correctness before being used. Literature identified with the assistance of generative AI was independently verified by consulting the original publications and relevant sources; claims and citations were not included solely on the basis of AI-generated output.

We take responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code accompanying this work will be made publicly available upon publication.

REFERENCES

Julie Alberge, Vincent Maladiere, Olivier Grisel, Judith Abécassis, and Gael Varoquaux. Survival Models: Proper Scoring Rule and Stochastic Optimization with Competing Risks. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pp. 3619–3627. PMLR, April 2025. URL https://proceedings.mlr.press/v258/alberge25a.html.

Guillaume Ausset, Tom Ciffreo, Francois Portier, Stephan Clémençon, and Timothée Papin. Individual Survival Curves with Conditional Normalizing Flows. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10, October 2021. doi: 10.1109/DSAA53316.2021.9564222. URL https://ieeexplore.ieee.org/abstract/document/9564222.

Anand Avati, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam H. Shah, and Andrew Y. Ng. Countdown Regression: Sharp and Calibrated Survival Predictions. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, pp. 145–155. PMLR, August 2020. URL https://proceedings.mlr.press/v115/avati20a.html.

Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence Carin Duke, and Ricardo Henao. Adversarial Time-to-Event Modeling. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 735–744. PMLR, July 2018. URL https://proceedings.mlr.press/v80/chapfuwa18a.html.

Ramon F. A. de Punder, Cees G. H. Diks, Roger J. A. Laeven, and Dick J. C. van Dijk. Localizing Strictly Proper Scoring Rules. *Journal of the American Statistical Association*, 0(0):1–13, January 2026. ISSN 0162-1459. doi: 10.1080/01621459.2025.2576189. URL https://doi.org/10.1080/01621459.2025.2576189.

Thomas A. Gerds and Martin Schumacher. Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal*, 48(6):1029–1040, 2006. ISSN 1521-4036. doi: 10.1002/bimj.200610301. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200610301.

Richard D. Gill, Mark J. van der Laan, and James M. Robins. Coarsening at Random: Characterizations, Conjectures, Counter-Examples. In D. Y. Lin and T. R. Fleming (eds.), *Proceedings of the First Seattle Symposium in Biostatistics*, pp. 255–294, New York, NY, 1997. Springer US. ISBN 978-1-4684-6316-3. doi: 10.1007/978-1-4684-6316-3_14.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN

0162-1459. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

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. ISSN 1097-0258. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%2819990915/30%2918%3A17/18%3C2529%3A%3AAID-SIM274%3E3.0.CO%3B2-5.

Xintian Han, Mark Goldstein, and Rajesh Ranganath. Survival Mixture Density Networks. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, pp. 224–248. PMLR, December 2022. URL https://proceedings.mlr.press/v182/han22a.html.

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, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. URL https://www.nature.com/articles/s41597-022-01899-x.

Jef Jonkers, Anahita Rouzé, Jarne Verhaeghe, Wim Van Biesen, Jan De Waele, Luc Duchateau, Glenn Van Wallendael, Eric Hoste, and Sofie Van Hoecke. Beyond binary AKI classification: development and external validation of a distributional model predicting serum creatinine and urine output trajectories in ICU patients. *Critical Care*, 30(1):249, May 2026. ISSN 1364-8535. doi: 10.1186/s13054-026-06017-6. URL https://doi.org/10.1186/s13054-026-06017-6.

Acute Kidney Injury Work Group KDIGO. KDIGO Clinical Practice Guideline for Acute Kidney Injury. *Kidney International Supplements*, 2(1):1–138, March 2012. ISSN 2157-1724. doi: 10.1038/kisup.2012.1. URL http://www.scopus.com/inward/record.url?scp=84884342549&partnerID=8YFLogxK.

Håvard Kvamme and Ørnulf Borgan. The Brier Score under Administrative Censoring: Problems and a Solution. *Journal of Machine Learning Research*, 24(2):1–26, 2023. ISSN 1533-7928. URL http://jmlr.org/papers/v24/19-1030.html.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468. doi: 10.1609/aaai.v32i1.11842. URL https://ojs.aaai.org/index.php/AAAI/article/view/11842.

Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep Survival Machines: Fully Parametric Survival Regression and Representation Learning for Censored Data With Competing Risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, August 2021. ISSN 2168-2208. doi: 10.1109/JBHI.2021.3052441. URL https://ieeexplore.ieee.org/abstract/document/9326348.

Alexander Norcliffe, Bogdan Cebere, Fergus Imrie, Pietro Lió, and Mihaela van der Schaar. SurvivalGAN: Generating Time-to-Event Data for Survival Analysis. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 10279–10304. PMLR, April 2023. URL https://proceedings.mlr.press/v206/norcliffe23a.html.

Tim Pearce, Jong-Hyeon Jeong, Yichen Jia, and Jun Zhu. Censored Quantile Regression Neural Networks for Distribution-Free Survival Analysis. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7450–7461. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/312c92e0cb862422eaa49452cdf55caf-Abstract-Conference.html.

Stephen Portnoy. Censored Regression Quantiles. *Journal of the American Statistical Association*, 98 (464):1001–1012, December 2003. ISSN 0162-1459. doi: 10.1198/016214503000000954. URL https://doi.org/10.1198/016214503000000954.

David Rindt, Robert Hu, David Steinsaltz, and Dino Sejdinovic. Survival regression with proper scoring rules and monotonic neural networks. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 1190–1205. PMLR, May 2022. URL https://proceedings.mlr.press/v151/rindt22a.html.

Xinwei Shen and Nicolai Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3): 653–677, July 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkae108. URL https://doi.org/10.1093/jrsssb/qkae108.

Hiroki Yanagisawa. Proper Scoring Rules for Survival Analysis. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 39165–39182. PMLR, July 2023. URL https://proceedings.mlr.press/v202/yanagisawa23a.html.

Hiroki Yanagisawa and Shunta Akiyama. A Strictly Proper Scoring Rule and a Calibration Metric for Interval-Censored Data Analysis. In *Forty-third International Conference on Machine Learning*, June 2026. URL https://openreview.net/forum?id=8jViL8YrB1.

A Theory and proofs

A.1 Proof of Proposition 1

We use $S_c^\flat(F;\psi_c^\flat(t))$ and $S_c^\flat(F;Y,\Delta)$ interchangeably for the abstract and right-censoring encodings of the same fixed-c censored observation.

Proof. Fix x. Let P_x denote the true conditional law of $T \mid X = x$ and let F_x denote a candidate forecast. Let $H(\cdot \mid x)$ be the true conditional distribution of the censoring time $C \mid X = x$. For each fixed censoring time c, define the induced censored laws

$$P_{c,x}^\flat := \mathcal{L}(\psi_c^\flat(T) \mid X = x), \qquad F_{c,x}^\flat := \mathcal{L}(\psi_c^\flat(Z) \mid X = x), \quad Z \sim F_x.$$

By definition,

$$\bar{\mathcal{S}}_x(F) - \bar{\mathcal{S}}_x(P) = \mathbb{E}_P\left[\bar{S}(F;Y,\Delta,X) - \bar{S}(P;Y,\Delta,X) \mid X = x\right].$$

The marginalized score averages the fixed-c localized score over the conditional law of C given the observed data,

$$\Pi(dc \mid Y,\Delta,X) := \mathcal{L}(C \in dc \mid Y,\Delta,X),$$

where this conditional law is induced by the true censoring distribution $H(\cdot \mid X)$. Hence, by the tower property,

$$\begin{aligned}\bar{\mathcal{S}}_x(F) - \bar{\mathcal{S}}_x(P) &= \mathbb{E}_P\Big[\mathbb{E}_P\Big[S_C^\flat(F_x;Y,\Delta) - S_C^\flat(P_x;Y,\Delta) \mid Y,\Delta,X\Big]\Big|X = x\Big] \\ &= \mathbb{E}_P\Big[S_C^\flat(F_x;Y,\Delta) - S_C^\flat(P_x;Y,\Delta) \mid X = x\Big].\end{aligned}$$

Under conditional independent censoring, $T \perp C \mid X$, we may condition on the realized censoring time and integrate over its true conditional law:

$$\bar{\mathcal{S}}_x(F) - \bar{\mathcal{S}}_x(P) = \int \mathbb{E}_{P_x}\Big[S_c^\flat(F_x;\psi_c^\flat(T)) - S_c^\flat(P_x;\psi_c^\flat(T))\Big]\, H(dc \mid x).$$

For fixed c, the distribution of $\psi_c^\flat(T)$ under the truth is exactly $P_{c,x}^\flat$. Therefore the inner expectation is the divergence of the localized score between the induced censored laws:

$$\mathbb{E}_{P_x}\Big[S_c^\flat(F_x;\psi_c^\flat(T)) - S_c^\flat(P_x;\psi_c^\flat(T))\Big] = D_{S_c^\flat}\Big(P_{c,x}^\flat \| F_{c,x}^\flat\Big).$$

Hence

$$\bar{\mathcal{S}}_x(F) - \bar{\mathcal{S}}_x(P) = \int D_{S_c^\flat}\Big(P_{c,x}^\flat \| F_{c,x}^\flat\Big)\, H(dc \mid x).$$

By propriety of the fixed-censoring localized score, the integrand is nonnegative for $H(\cdot \mid x)$-almost every c. Consequently,

$$\bar{\mathcal{S}}_x(F) \geq \bar{\mathcal{S}}_x(P).$$

If, in addition, the localized scores are strictly proper, then the integrand vanishes if and only if

$$P^\flat_{c,x} = F^\flat_{c,x}.$$

Since the integrand is nonnegative,

$$\bar{\mathcal{S}}_x(F) = \bar{\mathcal{S}}_x(P)$$

if and only if

$$P^\flat_{c,x} = F^\flat_{c,x}$$

for $H(\cdot \mid x)$-almost every c.

It remains to relate equality of the induced censored laws to the latent joint CDF. Fix $t = (t_1, \ldots, t_k) \in \mathcal{I}_x$.

By definition of $\mathcal{I}_x$, the set of censoring configurations satisfying $c_j > t_j$ for every j has positive $H(\cdot \mid x)$-measure (or, for shared censoring, $c > \max_j t_j$).

Since equality of the induced censored laws holds $H(\cdot \mid x)$-almost everywhere, there exists such a censoring configuration c for which $P^\flat_{c,x} = F^\flat_{c,x}$.

For this c, $\{\psi^\flat_c(T) \leq t\} = \{T \leq t\}$ where inequalities are interpreted componentwise. Hence equality of the induced censored laws implies

$$\mathbb{P}_{Z \sim F_x}(Z \leq t) = \mathbb{P}_{T \sim P_x}(T \leq t).$$

Since $t \in \mathcal{I}_x$ was arbitrary, the latent joint CDFs agree throughout $\mathcal{I}_x$. If $\mathcal{I}_x = \mathbb{R}^k_+$, equality of the joint CDFs implies equality of the full latent probability laws. □

A.2 Population optimality of censored engression

Proposition 2 (Population optimality of censored engression)**.** *Assume conditional independent censoring, $T \perp C \mid X$, and assume that $\Pi(dc \mid Y, \Delta, X)$ is the true forecast-independent conditional law of the censoring time given (Y, Δ, X). Assume that, for $H(\cdot \mid X = x)$-almost every censoring time c, the fixed-c censored energy score is well-defined and strictly proper for the corresponding censored law. Suppose there exists $g_0 \in \mathcal{M}$ such that*

$$g_0(x, \varepsilon) \sim P(T \mid X = x) \qquad \text{for } P_X\text{-almost every } x.$$

Let $\widetilde{g} \in \arg\min_{g \in \mathcal{M}} \mathcal{R}_{\text{cens-ES}}(g)$. Then

$$\mathcal{L}(\psi_c(\widetilde{g}(x, \varepsilon)) \mid X = x) = \mathcal{L}(\psi_c(T) \mid X = x)$$

for $H(\cdot \mid X = x)$-almost every c and P_X-almost every x. Consequently, the joint conditional CDF induced by $\widetilde{g}(x, \varepsilon)$ agrees with that of $T \mid X = x$ for every $t \in \mathcal{I}_x$. If $\mathcal{I}_x = \mathbb{R}^k_+$, the full conditional laws agree.

Proof. Fix x and a censoring time c. By assumption, the fixed-c censored energy score is strictly proper for the censored law $\mathcal{L}(\psi_c(T) \mid X = x)$. Hence, the corresponding fixed-c expected loss is minimized when

$$\mathcal{L}(\psi_c(g(x, \varepsilon)) \mid X = x) = \mathcal{L}(\psi_c(T) \mid X = x).$$

Since $\Pi(dc \mid Y, \Delta, X)$ is the true forecast-independent conditional law of the censoring time given the observed data, the population objective is the expectation of the corresponding observation-level marginalized score. Under correct specification, g_0 induces the true latent law and therefore attains zero integrated divergence, so it minimizes the population objective. Any other population minimizer $\widetilde{g}$ must also attain zero integrated divergence. Since the integrand is nonnegative for $H(\cdot \mid X = x)$-almost every c, it follows that

$$\mathcal{L}(\psi_c(\widetilde{g}(x, \varepsilon)) \mid X = x) = \mathcal{L}(\psi_c(T) \mid X = x)$$

for $H(\cdot \mid X = x)$-almost every c and P_X-almost every x. The joint-CDF identification claim follows by the same argument as in Proposition 1. □

B SPECIFIC SCORE DERIVATIONS

In each example, we proceed in the same way: we start from the uncensored score, define a fixed-censoring localized score, marginalize over the censoring time to obtain an observed-data score, and then relate this observed-data score back to the original latent target.

B.1 LOGARITHMIC SCORE

For the logarithmic score, the marginalization step simplifies substantially. Under right censoring, the fixed-c localized score depends on the censoring time only through the censored contribution, which is either independent of c or observed exactly from the data.

For fixed censoring time c, the induced censored law on the observed space has a density-mass representation

$$q_c^F(y, \delta \mid x) = \begin{cases} f(y \mid x), & \delta = 1,\ y \le c, \\ S(c \mid x), & \delta = 0,\ y = c, \end{cases}$$

where $f(\cdot \mid x)$ is the conditional density of T and $1 - F(t \mid x) = \mathbb{P}(T > t \mid X = x)$ is the corresponding survival function. Hence, the fixed-c localized logarithmic score is

$$S^\flat_{\log,c}(F; y, \delta, x) = -\log q_c^F(y, \delta \mid x) = -\delta \log f(y \mid x) - (1 - \delta) \log(1 - F(c \mid x)). \tag{4}$$

The marginalized score is

$$\bar{S}_{\log}(F; y, \delta, x) = \mathbb{E}\Big[S^\flat_{\log,C}(F; y, \delta, x) \mid Y = y, \Delta = \delta, X = x \Big].$$

We distinguish the two possible values of δ.

If $\delta = 1$, then conditioning on $Y = y, \Delta = 1, X = x$ implies that $T = y$ and $C \ge y$. By equation 4,

$$S^\flat_{\log,C}(F; y, 1, x) = -\log f(y \mid x),$$

which does not depend on C. Therefore, $\bar{S}_{\log}(F; y, 1, x) = -\log f(y \mid x)$.

If $\delta = 0$, then conditioning on $Y = y, \Delta = 0, X = x$ implies that $C = y$. Again by equation 4,

$$S^\flat_{\log,C}(F; y, 0, x) = -\log(1 - F(C \mid x)) = -\log(1 - F(y \mid x)),$$

and thus $\bar{S}_{\log}(F; y, 0, x) = -\log(1 - F(y \mid x))$.

Combining the two cases yields

$$\bar{S}_{\log}(F; y, \delta, x) = -\delta \log f(y \mid x) - (1 - \delta) \log(1 - F(y \mid x)).$$

In discrete or mixed settings, a boundary convention is needed at $Y = C$. With the convention $\Delta = \mathbb{1}\{T \le C\}$, the censored contribution is naturally written in terms of $\mathbb{P}(T > Y \mid X)$, i.e. $1 - F(Y \mid X)$.

B.2 CRPS

For the CRPS, the marginalization step is nontrivial. Under right censoring, the fixed-c localized score depends on the censoring time through the upper integration limit, so averaging over the conditional law of C is essential.

Recall that for a distribution function $F(\cdot \mid x)$ and observation t, the CRPS is

$$S_{\mathrm{CRPS}}(F; t, x) = \int_0^t F(s \mid x)^2 \, ds + \int_t^\infty (1 - F(s \mid x))^2 \, ds.$$

For fixed censoring time c, the censored law is the law of $\min(T, c)$, whose distribution function is

$$F_c^\flat(s \mid x) = \begin{cases} F(s \mid x), & s < c, \\ 1, & s \ge c. \end{cases}$$

Hence, the fixed-c localized CRPS is $S^{\flat}_{\mathrm{CRPS},c}(F;y,\delta,x) = S_{\mathrm{CRPS}}(F^{\flat}_c;y,x)$, which can be written as

$$S^{\flat}_{\mathrm{CRPS},c}(F;y,\delta,x) = \int_0^y F(s \mid x)^2\,ds + \delta \int_y^c (1-F(s \mid x))^2\,ds. \tag{5}$$

If $\delta = 1$, then $y \le c$ is observed exactly and the second term runs only up to the censoring time c, because $F^{\flat}_c(s \mid x) = 1$ for $s \ge c$. If $\delta = 0$, then necessarily $y = c$, and the second term vanishes.

The marginalized CRPS is

$$\bar{S}_{\mathrm{CRPS}}(F;y,\delta,x) = \mathbb{E}\Big[S^{\flat}_{\mathrm{CRPS},C}(F;y,\delta,x) \mid Y = y, \Delta = \delta, X = x\Big].$$

If $\delta = 0$, then conditioning on $Y = y, \Delta = 0, X = x$ implies that $C = y$, so

$$\bar{S}_{\mathrm{CRPS}}(F;y,0,x) = \int_0^y F(s \mid x)^2\,ds.$$

If $\delta = 1$, then conditioning on $Y = y, \Delta = 1, X = x$ implies that $C \ge y$. Using equation 5,

$$\bar{S}_{\mathrm{CRPS}}(F;y,1,x) = \int_0^y F(s \mid x)^2\,ds + \mathbb{E}\left[\int_y^C (1-F(s \mid x))^2\,ds \;\middle|\; C \ge y, X = x\right].$$

By Fubini's theorem,

$$\mathbb{E}\left[\int_y^C (1-F(s \mid x))^2\,ds \;\middle|\; C \ge y, X = x\right] = \int_y^\infty \mathbb{P}(C > s \mid C \ge y, X = x)\,(1-F(s \mid x))^2\,ds.$$

Writing $G(t^- \mid x) := \mathbb{P}(C \ge t \mid X = x)$, this becomes

$$\int_y^\infty \frac{G(s \mid x)}{G(y^- \mid x)}(1-F(s \mid x))^2\,ds.$$

Therefore,

$$\bar{S}_{\mathrm{CRPS}}(F;y,\delta,x) = \int_0^y F(s \mid x)^2\,ds + \delta \int_y^\infty \frac{G(s \mid x)}{G(y^- \mid x)}(1-F(s \mid x))^2\,ds. \tag{6}$$

This yields the marginalized right-censored CRPS induced by our construction.

B.3 Energy score

For the energy score, the fixed-censoring localized score is most naturally written in its expectation form. Unlike the logarithmic score, the marginalization step does not simplify away in general, because the fixed-censoring score depends on the censoring configuration through both expectation terms. In the univariate case with $\beta = 1$, the energy score reduces to the CRPS.

Let $\beta \in (0,2)$, and recall that for a predictive law $F(\cdot \mid x)$ on $\mathbb{R}^k_+$, the energy score is

$$S_{\mathrm{ES},\beta}(F;t,x) = \mathbb{E}\big[\|Z-t\|^\beta \mid X = x\big] - \frac{1}{2}\mathbb{E}\big[\|Z-Z'\|^\beta \mid X = x\big],$$

where $Z, Z' \overset{\text{iid}}{\sim} F(\cdot \mid x)$.

For fixed componentwise censoring threshold $c = (c_1,\dots,c_k)$, define

$$\psi_c(t) = (\min(t_1,c_1),\dots,\min(t_k,c_k)).$$

If $T \sim F(\cdot \mid X = x)$, the induced censored law is

$$F^{\flat}_c(\cdot \mid x) := \mathcal{L}(\psi_c(T) \mid X = x),$$

and the observed censored outcome is $y = \psi_c(t)$. Hence the fixed-c localized energy score is

$$S^{\flat}_{\mathrm{ES},\beta,c}(F; y, x) = \mathbb{E}[\|\psi_c(Z) - y\|^{\beta} \mid X = x] - \frac{1}{2}\mathbb{E}[\|\psi_c(Z) - \psi_c(Z')\|^{\beta} \mid X = x] \tag{7}$$

where $Z, Z' \overset{\text{iid}}{\sim} F(\cdot \mid X = x)$.

If the censoring threshold is random, the marginalized energy score is obtained by averaging equation 7 over the conditional law of C given the observed data:

$$\bar{S}_{\mathrm{ES},\beta}(F; Y, \Delta, X) = \mathbb{E}\Big[S^{\flat}_{\mathrm{ES},\beta,C}(F; Y, X) \mid Y, \Delta, X\Big].$$

In contrast to the logarithmic score and the CRPS, no further simplification is available in general, since the fixed-c localized score depends explicitly on c through both expectation terms.

B.4 Brier score

Let $\tau \geq 0$ denote a fixed evaluation horizon. The uncensored Brier score at horizon τ is

$$S_{\mathrm{BS},\tau}(F; t, x) = \big(F(\tau \mid x) - \mathbb{1}\{t \leq \tau\}\big)^2. \tag{8}$$

Note that we can do similar derivations for the binomial log-likelihood for $\tau \geq 0$.

For fixed censoring time c, let $Y_c := \min(T, c)$, $F^{\flat}_c := \mathcal{L}(Y_c \mid X = x)$. The natural localized Brier score at horizon τ is the ordinary Brier score applied to the censored distribution:

$$S^{\flat}_{\mathrm{BS},\tau,c}(F; t, x) := \big(F^{\flat}_c(\tau \mid x) - \mathbb{1}\{\min(t, c) \leq \tau\}\big)^2. \tag{9}$$

Since

$$F^{\flat}_c(\tau \mid x) = \mathbb{P}(\min(T, c) \leq \tau \mid X = x) = \begin{cases} F(\tau \mid x), & \tau < c, \\ 1, & \tau \geq c, \end{cases}$$

the score can be written more explicitly as

$$S^{\flat}_{\mathrm{BS},\tau,c}(F; t, x) = \begin{cases} \big(F(\tau \mid x) - \mathbb{1}\{t \leq \tau\}\big)^2, & \tau < c, \\ 0, & \tau \geq c. \end{cases} \tag{10}$$

Thus, for each fixed c, the localized Brier score coincides with the ordinary Brier score on the identifiable region $\tau < c$ and becomes vacuous beyond it. In particular, because the ordinary Brier score is strictly proper for Bernoulli outcomes, equation 9 is strictly proper for the censored Bernoulli target $\mathbb{1}\{\min(T, c) \leq \tau\}$, and therefore locally proper for the latent law on horizons $\tau < c$.

We now marginalize over the censoring time. Let $Y = \min(T, C)$, $\Delta = \mathbb{1}\{T \leq C\}$, and define

$$\bar{S}_{\mathrm{BS},\tau}(F; Y, \Delta, X) := \mathbb{E}\Big[S^{\flat}_{\mathrm{BS},\tau,C}(F; Y, X) \mid Y, \Delta, X\Big].$$

Using equation 10, we obtain the observed-data score by distinguishing three cases. If $Y > \tau$, then necessarily $C > \tau$ and $T > \tau$, so

$$\bar{S}_{\mathrm{BS},\tau}(F; Y, \Delta, X) = F(\tau \mid X)^2.$$

If $Y \leq \tau$ and $\Delta = 0$, then $C = Y \leq \tau$, hence $\tau \geq C$ and

$$\bar{S}_{\mathrm{BS},\tau}(F; Y, 0, X) = 0.$$

If $Y \leq \tau$ and $\Delta = 1$, then $T = Y \leq \tau$ and $C \geq Y$. In this case,

$$S^{\flat}_{\mathrm{BS},\tau,C}(F; Y, X) = \begin{cases} (1 - F(\tau \mid X))^2, & C > \tau, \\ 0, & Y \leq C \leq \tau. \end{cases}$$

Therefore,

$$\bar{S}_{\mathrm{BS},\tau}(F; Y, 1, X) = \mathbb{P}(C > \tau \mid C \geq Y, X)\,(1 - F(\tau \mid X))^2.$$

Writing

$$G(t \mid x) := \mathbb{P}(C > t \mid X = x), \qquad G(t^- \mid x) := \mathbb{P}(C \geq t \mid X = x),$$

this becomes

$$\bar{S}_{\mathrm{BS},\tau}(F; Y, 1, X) = \frac{G(\tau \mid X)}{G(Y^- \mid X)}(1 - F(\tau \mid X))^2.$$

Combining the three cases yields

$$\bar{S}_{\mathrm{BS},\tau}(F; Y, \Delta, X) = \mathbb{1}\{Y > \tau\}F(\tau \mid X)^2 + \Delta\,\mathbb{1}\{Y \leq \tau\}\frac{G(\tau \mid X)}{G(Y^- \mid X)}(1 - F(\tau \mid X))^2. \quad (11)$$

This is the observed-data version of the localized proper Brier score.

A convenient rescaling recovers the usual IPCW Brier score: dividing equation 11 by $G(\tau \mid X)$ yields

$$\frac{\bar{S}_{\mathrm{BS},\tau}(F; Y, \Delta, X)}{G(\tau \mid X)} = \frac{\Delta\mathbb{1}\{Y \leq \tau\}(1 - F(\tau \mid X))^2}{G(Y^- \mid X)} + \frac{\mathbb{1}\{Y > \tau\}F(\tau \mid X)^2}{G(\tau \mid X)},$$

which is exactly the standard IPCW Brier score (Graf et al., 1999; Gerds & Schumacher, 2006). Since this rescales the score by a strictly positive factor that does not depend on F, it preserves the same minimizer and hence the same propriety characteristics on the identifiable region.

B.4.1 Integrated Brier score

The integrated version fits especially naturally into the localization framework. Define the fixed-c localized integrated Brier score by

$$S^{\flat}_{\mathrm{IBS},c}(F; t, x) := \int_0^\infty S^{\flat}_{\mathrm{BS},\tau,c}(F; t, x)\, d\tau. \quad (12)$$

Using equation 10, this reduces to

$$S^{\flat}_{\mathrm{IBS},c}(F; t, x) = \int_0^c \big(F(\tau \mid x) - \mathbb{1}\{t \leq \tau\}\big)^2\, d\tau. \quad (13)$$

Thus the fixed-censoring localized IBS is simply the ordinary integrated Brier score truncated to the identifiable region $[0, c)$.

Marginalizing over the censoring time gives

$$\bar{S}_{\mathrm{IBS}}(F; Y, \Delta, X) = \mathbb{E}\Big[S^{\flat}_{\mathrm{IBS},C}(F; Y, X) \mid Y, \Delta, X\Big].$$

By Fubini's theorem and equation 11,

$$\begin{aligned}\bar{S}_{\mathrm{IBS}}(F; Y, \Delta, X) &= \int_0^\infty \bar{S}_{\mathrm{BS},\tau}(F; Y, \Delta, X)\, d\tau \\ &= \int_0^Y F(\tau \mid X)^2\, d\tau + \Delta \int_Y^\infty \frac{G(\tau \mid X)}{G(Y^- \mid X)}(1 - F(\tau \mid X))^2\, d\tau. \quad (14)\end{aligned}$$

This is the integrated observed-data score induced by the localized proper Brier construction.

In particular, equation 14 has the same structural form as the marginalized CRPS: an integral of squared CDF error up to the observed time, plus a tail term weighted by the conditional survival of the censoring time. In the univariate continuous setting, this is not merely a formal similarity: the integrated Brier score and the CRPS coincide, so the marginalized IBS and the marginalized CRPS are the same score up to the usual boundary convention.

B.5 Pinball loss

Let $\alpha \in (0, 1)$ denote a quantile level, and let

$$q_\alpha(F \mid x) := \inf\{t \in \mathbb{R}_+ : F(t \mid x) \geq \alpha\}$$

be the corresponding conditional α-quantile implied by the forecast $F(\cdot \mid x)$. The uncensored quantile score is

$$S_{\mathrm{Q},\alpha}(F;t,x) = \big(\alpha - \mathbb{1}\{t < q_\alpha(F \mid x)\}\big)\big(t - q_\alpha(F \mid x)\big). \tag{15}$$

Equivalently,

$$S_{\mathrm{Q},\alpha}(F;t,x) = \alpha\big(t - q_\alpha(F \mid x)\big)_+ + (1-\alpha)\big(q_\alpha(F \mid x) - t\big)_+.$$

This score is strictly consistent for the α-quantile functional.

For fixed censoring time c, let $Y_c := \min(T,c)$, $F_c^\flat := \mathcal{L}(Y_c \mid X = x)$. The natural localized quantile score is the ordinary quantile score applied to the censored law:

$$S^\flat_{\mathrm{Q},\alpha,c}(F;t,x) := \big(\alpha - \mathbb{1}\{\min(t,c) < q_\alpha(F_c^\flat \mid x)\}\big)\big(\min(t,c) - q_\alpha(F_c^\flat \mid x)\big). \tag{16}$$

Since

$$q_\alpha(F_c^\flat \mid x) = \min\big(q_\alpha(F \mid x), c\big),$$

this can be written as

$$S^\flat_{\mathrm{Q},\alpha,c}(F;t,x) = \big(\alpha - \mathbb{1}\{\min(t,c) < \min(q_\alpha(F \mid x), c)\}\big)\big(\min(t,c) - \min(q_\alpha(F \mid x), c)\big). \tag{17}$$

Thus the fixed-censoring localized score is simply the ordinary quantile score for the censored outcome $\min(T,c)$ and the censored quantile $\min(q_\alpha(F \mid x), c)$.

Let $Y = \min(T,C)$, $\Delta = \mathbb{1}\{T \le C\}$, and define

$$\bar{S}_{\mathrm{Q},\alpha}(F;Y,\Delta,X) := \mathbb{E}\Big[S^\flat_{\mathrm{Q},\alpha,C}(F;Y,X) \mid Y,\Delta,X\Big].$$

Write

$$q_\alpha(X) := q_\alpha(F \mid X), \qquad G(t \mid x) := \mathbb{P}(C > t \mid X = x), \qquad G(t^- \mid x) := \mathbb{P}(C \ge t \mid X = x).$$

If $\Delta = 0$, then $C = Y$, so $q_\alpha(F_C^\flat \mid X) = \min(q_\alpha(F \mid X), Y)$, and therefore

$$\bar{S}_{\mathrm{Q},\alpha}(F;Y,0,X) = \big(\alpha - \mathbb{1}\{Y < \min(q_\alpha(F \mid X), Y)\}\big)\big(Y - \min(q_\alpha(F \mid X), Y)\big) = \alpha\big(Y - q_\alpha(F \mid X)\big)_+.$$

If $\Delta = 1$, then $T = Y$ and $C \ge Y$. In this case,

$$\bar{S}_{\mathrm{Q},\alpha}(F;Y,1,X) = \mathbb{E}\big[\big(\alpha - \mathbb{1}\{Y < \min(q_\alpha(F \mid X), C)\}\big)\big(Y - \min(q_\alpha(F \mid X), C)\big) \mid C \ge Y, X\big].$$

If $q_\alpha(F \mid X) \le Y$, then $\min(q_\alpha(F \mid X), C) = q_\alpha(F \mid X)$ for every $C \ge Y$, so

$$\bar{S}_{\mathrm{Q},\alpha}(F;Y,1,X) = \alpha\big(Y - q_\alpha(F \mid X)\big).$$

If $q_\alpha(F \mid X) > Y$, then

$$\big(\alpha - \mathbb{1}\{Y < \min(q_\alpha(F \mid X), C)\}\big)\big(Y - \min(q_\alpha(F \mid X), C)\big) = (1-\alpha)\big(\min(q_\alpha(F \mid X), C) - Y\big),$$

and hence

$$\begin{aligned}\bar{S}_{\mathrm{Q},\alpha}(F;Y,1,X) &= \frac{1-\alpha}{G(Y^- \mid X)}\mathbb{E}\big[\big(\min(q_\alpha(F \mid X), C) - Y\big)\mathbb{1}\{C \ge Y\} \mid X\big] \\ &= \frac{1-\alpha}{G(Y^- \mid X)}\int_Y^{q_\alpha(F|X)} G(t \mid X)\,dt.\end{aligned}$$

Combining the two cases yields

$$\bar{S}_{\mathrm{Q},\alpha}(F;Y,\Delta,X) = \alpha\big(Y - q_\alpha(F \mid X)\big)_+ + \Delta\,\frac{1-\alpha}{G(Y^- \mid X)}\int_Y^{q_\alpha(F|X)} G(t \mid X)\,dt, \tag{18}$$

with the convention that the integral is zero whenever $q_\alpha(X) \le Y$.

Table 4: Computational diagnostics for the synthetic log-normal censored-engression experiments. Censoring rate, training time, and epochs are mean ± standard deviation over five repetitions. Peak memory is the largest per-process resident-set size observed across the five runs.

k	Censoring	Censored (%)	Time (s)	Peak MiB	Epochs
2	Administrative	15.9 ± 0.4	49.0 ± 5.6	457.5	31.0 ± 3.7
2	Uniform	35.7 ± 0.2	51.5 ± 6.7	462.5	35.0 ± 5.2
2	Conditional uniform	43.1 ± 0.3	46.9 ± 6.4	463.3	32.0 ± 4.9
5	Administrative	17.8 ± 0.2	51.3 ± 5.1	458.2	30.0 ± 3.5
5	Uniform	36.5 ± 0.3	73.8 ± 7.4	458.1	45.0 ± 5.0
5	Conditional uniform	43.7 ± 0.2	66.2 ± 8.0	458.2	39.0 ± 5.1
10	Administrative	22.7 ± 0.1	92.7 ± 10.2	445.5	45.0 ± 5.2
10	Uniform	38.7 ± 0.2	106.7 ± 16.2	472.5	52.0 ± 8.7
10	Conditional uniform	45.2 ± 0.2	91.2 ± 5.5	465.3	44.0 ± 2.9

C Additional experimental details and results

C.1 Computational cost and stability

The synthetic censored-engression experiments were run on an Apple M1 Pro with 32 GB unified memory using the MPS backend. Table 4 reports wall-clock training time, peak process memory, and epochs to early stopping over five independent repetitions.

All runs reached the early-stopping criterion. We observed no fitting errors, non-finite training losses, or non-finite validation-selection metrics. Training remained below two minutes per run in every synthetic setting. Runtime increased with event dimension, whereas peak process memory remained between approximately 445 and 473 MiB. Heavier censoring reduced statistical information and sometimes increased the number of epochs required, but did not produce optimization failures.

For the ICU experiment, training used 8 CPU cores of an Intel Xeon Platinum 8168 at 2.70 GHz, 64 GB system memory, and one NVIDIA Tesla V100-SXM3 GPU with 32 GB memory. Final model training required 8,392 seconds (139.9 minutes). Peak process resident-set size was 33,154 MiB and peak CUDA-reserved memory was 490 MiB. Training times include optimization and validation passes but exclude preprocessing and final test scoring.

C.2 Forecast-evaluation simulation setup

This subsection gives the exact simulation design for the forecast-evaluation experiment reported in Section 5.1. It specifies the latent event-time law, candidate forecasts, censoring regimes, discretization grid, and scoring rules used in the comparison.

Latent event-time model and grid. We simulate covariates $X = (X_1, X_2, X_3) \sim N(0, I_3)$ and latent event times from

$$T \mid X = x \sim \text{Weibull}(k, \lambda(x)), \qquad k = 1.5, \qquad \log \lambda(x) = 0.3 + 0.8x_1 - 0.5x_2 + 0.3x_3.$$

All reported results are based on a test sample of size $N = 1000$. To match the discrete forecast-based scores, we represent forecasts on a grid

$$0 = \zeta_0 < \zeta_1 < \cdots < \zeta_B = z_{\max}, \qquad B = 50,$$

where $z_{\max} = 20.5471$ is the empirical 0.995-quantile of a large pilot sample from the oracle distribution. We also evaluate continuous versions of our CRPS, pinball, Brier, and log scores before discretization when a continuous forecast representation is available; however, for conciseness, we omit them from Table 2, as they have the same ranking as their discrete counterparts. Note that their scores are lower, since the DGP is continuous and the discretization is an approximation.

Censoring regimes. We consider four regimes.

Regime A: administrative censoring. The censoring time is deterministic, $C \equiv c_{\text{admin}}$, with $c_{\text{admin}} = 0.9833$ in the realized simulation, giving an event rate of 50.00%. Because the censoring time is known exactly, we evaluate this regime using localized scores.

Regime B: random independent censoring. The censoring time is sampled independently of T and X:

$$C \sim \mathrm{Uniform}(0, C_{\max}), \qquad C_{\max} = 0.4 z_{\max} = 8.2188.$$

The event rate is 79.70%. Since the censoring law is marginal and forecast-independent, we evaluate this regime with marginalized scores using the true marginal censoring distribution.

Regime C: conditionally independent censoring. The censoring time depends on covariates but remains conditionally independent of T:

$$C \mid X = x \sim \mathrm{Weibull}(k, \mu_C(x)), \qquad k = 1.5, \qquad \log \mu_C(x) = 0.2 - 0.3x_1 + 0.4x_3.$$

The event rate is 48.60%. We evaluate this regime using marginalized scores with the true conditional censoring law.

Regime D: targeted forecast-dependent weights stress test. Let $j = \lfloor B/2 \rfloor = 25$, so that $\zeta_{j-1} = 9.8626$ and $\zeta_j = 10.2736$. Define

$$a_j = \zeta_{j-1} + 0.25(\zeta_j - \zeta_{j-1}) = 9.9653, \qquad b_j = \zeta_{j-1} + 0.50(\zeta_j - \zeta_{j-1}) = 10.0681.$$

The oracle event distribution places probability $\frac{1}{2}$ on $(b_j, \zeta_j]$ and probability $\frac{1}{2}$ on $(\zeta_{B-1}, \zeta_B]$, where $\zeta_{B-1} = 20.1362$ and $\zeta_B = 20.5471$. Censoring is generated by a mixture: with probability 0.6, $C \sim \mathrm{Uniform}(\zeta_{j-1}, a_j)$, so censoring occurs before the interior event mass; with probability 0.4, C is set to a late administrative value beyond $z_{\max}$, so the event is observed. The resulting event rate is 39.50%. This regime is designed to allow forecast-dependent completion to be gamed by concentrating mass in the far tail.

Candidate forecasts. In regimes A–C, we compare five forecasts. F_0 is the oracle Weibull forecast. F_1 is a shifted Weibull forecast with $\lambda_1(x) = e^{0.25}\lambda(x)$. F_2 is an overdispersed and right-shifted discrete forecast obtained by smoothing the oracle pmf with a symmetric kernel proportional to $(1, 2, 3, 2, 1)$ and shifting mass two bins to the right. F_3 is underdispersed and right-shifted, constructed by blending the oracle pmf with a narrow peaked template centered two bins to the right of the oracle median bin, using width 1.25 and mixing weight 0.70. F_4 is upper-tail-heavy, with $p_{4,i}(x) \propto p_{0,i}(x) \exp(i/B)$.

Regime D uses a targeted stress-test construction. There, the oracle $F_0^{(D)}$ places half its mass in an interior bin interval and half its mass in the last grid bin. It is compared with exploit forecasts $F_5(\epsilon)$, with $\epsilon \in \{10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}\}$, that place only ϵ mass in the interior event interval and nearly all remaining mass in the final bin, up to a small numerical probability floor.

Scoring rules and evaluation protocol. As infeasible gold standards, we evaluate latent CRPS, pinball, Brier, and negative log-likelihood scores with respect to the unobserved true event times. On censored data, we evaluate the forecast-independent scores proposed in this paper: localized scores in regime A and marginalized scores in regimes B–D.

For direct comparison with Yanagisawa (2023), we thus evaluate their right-censored forecast-dependent weighted log score, multiclass Brier score, binary Brier score, pinball score, and ranked probability score (RPS). Each is evaluated either with oracle weights computed from F_0, or with plug-in weights computed from the forecast being scored. The binary Brier score is averaged over event-time percentiles, using the targeted grid threshold in regime D.

C.3 Oracle rankings and raw score values

The main text summarizes the principal finding that forecast-independent scores preserve the oracle ordering, whereas forecast-dependent plug-in scores can induce ranking reversals. Here we report the complete oracle rankings and corresponding raw score values.

Table 2 reports the rank of the oracle forecast for each scoring rule across regimes A–D. The latent gold-standard scores rank the oracle first in every regime, as do all of our localized and marginalized scores. The oracle-weight versions of the Yanagisawa (2023) scores also rank the oracle first, confirming that these scores behave as intended when the required weights are evaluated using the correct event-time law.

Table 5: Raw mean scores for regimes A–C. Each entry gives the oracle forecast score, with the best competing forecast score in parentheses. Lower is better. Bold entries indicate ranking reversals, where the best competing forecast obtains a lower score than the oracle.

Scoring rule	A	B	C
CRPS, latent	0.7766 (0.7905)	0.7766 (0.7905)	0.7766 (0.7905)
Pinball, latent	0.3030 (0.3057)	0.3030 (0.3057)	0.3030 (0.3057)
Brier, latent	0.1016 (0.1021)	0.1016 (0.1021)	0.1016 (0.1021)
NLL, latent	1.0745 (1.1389)	1.0745 (1.1389)	1.0745 (1.1389)
NLL (Yanagisawa, 2023), plug-in	**0.9841 (0.9800)**	1.5163 (1.5171)	**0.9086 (0.9056)**
Brier MC (Yanagisawa, 2023), plug-in	0.7721 (0.7735)	0.7689 (0.7694)	0.7676 (0.7690)
Brier (Yanagisawa, 2023), plug-in	0.1041 (0.1045)	0.1034 (0.1040)	**0.1020 (0.1019)**
Pinball (Yanagisawa, 2023), plug-in	0.2947 (0.3400)	0.3286 (0.3477)	0.3935 (0.4197)
RPS (Yanagisawa, 2023), plug-in	**1.6839 (1.4671)**	1.6877 (1.7565)	**1.6683 (1.4206)**
NLL (Yanagisawa, 2023), oracle	0.9841 (0.9856)	1.5163 (1.5193)	0.9086 (0.9099)
Brier MC (Yanagisawa, 2023), oracle	0.7721 (0.7726)	0.7689 (0.7695)	0.7676 (0.7680)
Brier (Yanagisawa, 2023), oracle	0.1041 (0.1045)	0.1034 (0.1039)	0.1020 (0.1023)
Pinball (Yanagisawa, 2023), oracle	0.2947 (0.3400)	0.3286 (0.3448)	0.3935 (0.4178)
RPS (Yanagisawa, 2023), oracle	1.6839 (1.7096)	1.6877 (1.7098)	1.6683 (1.6938)
Censored NLL	0.3811 (0.4036)	0.7350 (0.7702)	0.3445 (0.3688)
CRPS, ours	0.1129 (0.1133)	0.3729 (0.3753)	0.1268 (0.1278)
Pinball, ours	0.0416 (0.0417)	0.1418 (0.1429)	0.0473 (0.0476)
Brier, ours	0.0457 (0.0460)	0.0873 (0.0880)	0.0494 (0.0497)

Table 6: Raw mean scores for regime D. Entries report the oracle score and the best exploit forecast score. Lower is better. Bold entries indicate ranking reversals, where an exploit forecast obtains a lower score than the oracle.

Scoring rule	Oracle score	Best exploit score
CRPS, latent	2.6189	4.8200
Pinball, latent	0.9320	1.8684
Brier, latent	0.0250	0.0455
NLL, latent	−0.5448	0.2944
NLL (Yanagisawa, 2023), plug-in	**0.6931**	**0.6716**
Brier MC (Yanagisawa, 2023), plug-in	**0.5000**	**0.3683**
Brier (Yanagisawa, 2023), plug-in	**0.0250**	**0.0184**
Pinball (Yanagisawa, 2023), plug-in	1.1289	1.4673
RPS (Yanagisawa, 2023), plug-in	**6.2500**	**4.6039**
NLL (Yanagisawa, 2023), oracle	0.6931	1.4235
Brier MC (Yanagisawa, 2023), oracle	0.5000	0.8438
Brier (Yanagisawa, 2023), oracle	0.0250	0.0422
Pinball (Yanagisawa, 2023), oracle	1.1289	1.5317
RPS (Yanagisawa, 2023), oracle	6.2500	10.5473
Censored NLL	−0.2050	0.0832
CRPS, ours	1.0195	1.6975
Pinball, ours	0.3741	0.6843
Brier, ours	0.0099	0.0167

The failures arise when those weights are recomputed from the candidate forecast. The plug-in NLL ranks the oracle second in regimes A, C, and D; the plug-in RPS ranks it second in A and C and last in D; and the plug-in Brier variants also exhibit reversals in C or D. In contrast, the plug-in pinball score preserves the oracle ordering in these experiments.

Tables 5 and 6 report the corresponding raw mean scores. They show that the ranking reversals result from competing or exploit forecasts attaining lower plug-in scores than the oracle, whereas the forecast-independent scores continue to assign the oracle the lowest score.

C.4 Sensitivity to estimating the censoring distribution

We use the same synthetic setup as in Section 5.1, with the exact simulation details given in Appendix C.2, but now replace the true censoring law in regimes B and C by estimated censoring distributions. The estimators are fit by treating censoring as the event. We consider a Kaplan–Meier estimator and a pooled Weibull model. In regime B, these pooled estimators are well matched to the DGP because censoring is independent of both T and X. In regime C, they are intentionally misspecified because they ignore the covariate dependence of $C \mid X$.

Table 7 reports the resulting raw score values for our marginalized CRPS, pinball, and Brier scores, using both discrete and continuous implementations where available. The rank summary is as follows. In regime B, both the Kaplan–Meier and pooled Weibull censoring estimators preserve the oracle ordering for all three discrete scores. In regime C, the pooled Kaplan–Meier estimator also ranks the oracle first for all three discrete scores. The pooled Weibull estimator preserves the oracle ranking for discrete CRPS and pinball, but the discrete Brier score ranks the near-oracle tail forecast slightly ahead of the oracle. This reflects ordinary censoring-model misspecification rather than forecast-dependent circularity: once the censoring estimator is fixed, all candidate forecasts are still evaluated by the same forecast-independent score.

C.5 Synthetic censored-engression setup

This subsection gives the DGPs and implementation details for the synthetic multivariate censored-engression experiments in Section 5.2. The experiments use simulated data so that the latent event times and the true data-generating conditional law are available for oracle evaluation.

Covariates. For each observation i, we draw

$$X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^\top \sim N(0, I_4).$$

The latent event-time vector has dimension k, where $k = 2$ for the bivariate experiments and $k \in \{3, 5, 10\}$ for the scaling experiments.

Table 7: Sensitivity of our scores to estimated censoring distributions. Entries are mean scores; lower is better. Regime B uses censoring independent of both T and X. Regime C uses covariate-dependent censoring, but the censoring estimator is pooled and therefore ignores the covariate dependence.

Regime / estimator	Forecast	CRPS disc.	CRPS cont.	Pinball disc.	Pinball cont.	Brier disc.	Brier cont.
Regime B, KM censoring estimator							
	F_0 oracle	0.4326	0.3748	0.1637	0.1420	0.1023	0.0875
	F_1 shifted	0.4863	0.4002	0.1827	0.1524	0.1093	0.0926
	F_2 over	0.7960	0.6497	0.2943	0.2427	0.2072	0.1819
	F_3 under	0.6281	0.5095	0.2346	0.1931	0.1470	0.1269
	F_4 tail	0.4405	0.3766	0.1666	0.1432	0.1031	0.0881
Regime B, Weibull censoring estimator							
	F_0 oracle	0.4326	0.3752	0.1640	0.1424	0.1028	0.0872
	F_1 shifted	0.4874	0.4037	0.1837	0.1535	0.1096	0.0922
	F_2 over	0.7926	0.6456	0.2935	0.2416	0.2070	0.1808
	F_3 under	0.6248	0.5061	0.2333	0.1919	0.1472	0.1261
	F_4 tail	0.4416	0.3786	0.1674	0.1441	0.1035	0.0878
Regime C, KM censoring estimator, pooled							
	F_0 oracle	0.2402	0.1295	0.0891	0.0482	0.1023	0.0496
	F_1 shifted	0.2618	0.1365	0.0965	0.0510	0.1084	0.0516
	F_2 over	0.4316	0.2708	0.1578	0.0998	0.2099	0.1073
	F_3 under	0.3301	0.1845	0.1198	0.0679	0.1466	0.0716
	F_4 tail	0.2432	0.1303	0.0901	0.0486	0.1028	0.0499
Regime C, Weibull censoring estimator, pooled							
	F_0 oracle	0.2399	0.1281	0.0888	0.0474	0.1279	0.0496
	F_1 shifted	0.2604	0.1334	0.0956	0.0495	0.1285	0.0517
	F_2 over	0.4352	0.2701	0.1587	0.0991	0.2300	0.1080
	F_3 under	0.3317	0.1838	0.1200	0.0674	0.1772	0.0720
	F_4 tail	0.2424	0.1288	0.0896	0.0476	0.1273	0.0499

Base coefficient functions. For both latent DGPs, we use the same covariate-dependent location and scale functions. Let

$$b = (-0.6, -0.2, 0.2, 0.6)^\top.$$

For event coordinate $j = 1, \dots, k$, define

$$\mu_j(x) = 0.35 + x^\top \beta_j, \qquad \sigma_j(x) = 0.35 + 0.15\,\mathrm{sigmoid}(x^\top \gamma_j),$$

where

$$\beta_j = (0.35 + 0.08(j-1))\big(\cos(b_1 + 0.7(j-1)), \dots, \cos(b_4 + 0.7(j-1))\big)^\top,$$

and

$$\gamma_j = 0.10\big(\sin(b_1 - 0.7(j-1)), \dots, \sin(b_4 - 0.7(j-1))\big)^\top.$$

Unimodal log-normal DGP. The original, easier DGP is a correlated log-normal model. Let $\varepsilon_i \sim N(0, \Sigma_\rho)$, where Σ_ρ is the $k \times k$ equicorrelation matrix with diagonal entries 1 and off-diagonal entries $\rho = 0.45$. Conditional on $X_i = x$, the latent log-times are

$$\log T_{ij} = \mu_j(x) + \sigma_j(x)\varepsilon_{ij}, \qquad j = 1, \dots, k.$$

We use this DGP only as a sanity check in the appendix, with $k = 2$.

Mixture log-normal DGP. The main experiments use a covariate-dependent two-regime mixture log-normal DGP. Let

$$\pi(x) = \mathrm{sigmoid}\left(0.8x_1 - 0.6x_2 + 0.4\sin(x_3) + 0.25(x_4^2 - 1)\right),$$

and draw a latent regime indicator

$$Z_i \mid X_i = x \sim \mathrm{Bernoulli}(\pi(x)).$$

The two regimes correspond to an early and a late event-time mode. Let ℓ_j be event-coordinate-specific loadings, evenly spaced between 0.85 and 1.15:

$$\ell_j = 0.85 + \frac{j-1}{k-1}(1.15 - 0.85), \qquad j = 1, \dots, k,$$

with $\ell_1 = 1$ when $k = 1$. The regime shift is

$$a_j = 1.15\,\ell_j.$$

Let $\varepsilon_i^{(0)} \sim N(0, \Sigma_0)$ and $\varepsilon_i^{(1)} \sim N(0, \Sigma_1)$, independently, where Σ_0 and Σ_1 are $k \times k$ equicorrelation matrices with off-diagonal correlations 0.70 and 0.20, respectively. The early regime has stronger dependence and smaller marginal scale, while the late regime has weaker dependence and larger marginal scale:

$$\log T_{ij}^{(0)} = \mu_j(x) - a_j + 0.85\,\sigma_j(x)\varepsilon_{ij}^{(0)},$$
$$\log T_{ij}^{(1)} = \mu_j(x) + a_j + 1.15\,\sigma_j(x)\varepsilon_{ij}^{(1)}.$$

The latent event time is then

$$\log T_{ij} = (1 - Z_i)\log T_{ij}^{(0)} + Z_i \log T_{ij}^{(1)}.$$

This DGP induces a nonlinear, covariate-dependent, multimodal conditional distribution and is intentionally outside the Weibull-copula family used by the parametric likelihood baselines.

Censoring mechanisms. A single scalar censoring time C_i is shared by all event coordinates. The observed outcomes are

$$Y_{ij} = \min(T_{ij}, C_i), \qquad \Delta_{ij} = \mathbb{1}\{T_{ij} \le C_i\}, \qquad j = 1, \dots, k.$$

Thus, censoring is perfectly dependent across coordinates through the shared censoring variable C_i.

We consider three censoring mechanisms:

- *Administrative censoring:*
$$C \equiv c_{\text{admin}} = 3.0.$$
- *Random independent censoring:*
$$C \sim \text{Uniform}(0, c_{\max}), \qquad c_{\max} = 5.0.$$
- *Conditionally independent censoring:*
$$C \mid X = x \sim \text{Uniform}(0, c_{\max}(x)),$$
with
$$c_{\max}(x) = 5.0\,(0.45 + 0.70\,\text{sigmoid}(0.45x_1 - 0.35x_2 + 0.20x_3)).$$

The administrative regime corresponds to fixed localized censoring. The two random censoring regimes require the marginalized censored-energy objective because, for uncensored observations, the realized censoring time is only known to exceed the largest observed event time.

Methods and baselines. We compare censored engression with oracle references, naive observed engression, and likelihood-based survival baselines. The *DGP* row evaluates samples from the true conditional law and is not a fitted method. *Latent engression* is an infeasible benchmark trained directly on latent event times T. *Naive observed engression* trains the same generator architecture on Y as if it were fully observed. *Censored engression* trains the generator with the localized or marginalized censored energy score.

The likelihood-based baselines are fitted only on (X, Y, Δ) using the censored logarithmic score. We include independent Weibull, conditional MLP independent Weibull, Clayton copula Weibull, and conditional MLP Clayton copula Weibull baselines. For $k = 2$, we additionally include Gumbel copula Weibull and joint discrete-grid likelihood baselines. The latter is included only in low dimension because its output size scales as $O(B^k)$ for B grid bins.

The conditional Weibull and Clayton-copula baselines use neural networks to parameterize the conditional distribution, so these comparisons are not simply between a neural method and fixed parametric models. Rather, they compare an implicit sample-based representation with neural likelihood models constrained to prespecified Weibull and copula families.

Training setup. All engression models use the same noise-conditioned MLP generator with hidden width 128, noise dimension 128, depth 2, and a softplus output transform. Training, validation, and test sizes are 4000, 1000, and 1000, respectively. We report means and standard errors over five independent repetitions.

Engression models are trained for at most 200 epochs with early stopping. For censored engression, we use 16 latent samples per minibatch update. For evaluation, we use 1024 latent samples. In the random censoring regimes, the marginalized score is approximated with 8 censoring draws during training and 512 censoring draws during evaluation.

The likelihood baselines are trained by minimizing their observed-data censored negative log-likelihood. Validation negative log-likelihood is used for early stopping. For the conditional MLP Weibull and conditional MLP Clayton baselines, we use hidden width 128, depth 2, batch size 512, and AdamW optimization.

C.6 Synthetic censored-engression results

Across the mixture log-normal experiments, censored engression achieves the lowest censored energy score among feasible methods across the reported censoring mechanisms and dimensions. The improvement over naive observed engression becomes particularly pronounced under random and covariate-dependent censoring, where treating the censored vector as fully observed produces a distorted observed-data distribution. For example, under uniform censoring the censored energy score decreases from 0.5524 to 0.4904 at $k = 2$ and from 1.6402 to 1.3382 at $k = 10$.

The likelihood-based baselines provide substantially stronger comparisons than naive observed training because they are fitted with the proper censored logarithmic score. The conditional MLP versions additionally allow the Weibull marginals and copula parameters to depend flexibly on the covariates, so the comparison is not simply neural versus parametric. At $k = 2$ under uniform censoring, for example, the MLP Clayton–Weibull model obtains censored ES 0.5152, compared with 0.4904 for censored engression. The remaining gap is consistent with the mixture log-normal DGP lying outside the prespecified Weibull-copula family, whereas censored engression represents the conditional joint distribution implicitly through generated samples.

The latent energy score serves a different purpose. It evaluates recovery of the full latent event-time law, including regions that are only partially identified from the censored observations. Likelihood-based copula models can therefore obtain lower latent energy score than censored engression in some settings even when they have worse censored energy score. This does not contradict the observed-data comparison: censored engression is optimized for the proper censored-data criterion, while behavior beyond the identifiable region is not fully determined by the observed censored distribution.

Table 12 reports the corresponding $k = 2$ results under the unimodal log-normal sanity-check DGP. This setting is more favorable to the Weibull-copula likelihood models. As expected, these models are highly competitive and can outperform censored engression on latent energy score in some settings, while censored engression remains competitive on censored energy score and continues to improve substantially over naive observed engression.

Table 8: Synthetic multivariate survival experiment under a covariate-dependent mixture log-normal latent DGP with $k = 2$ event times and one shared censoring time. Entries are mean $\pm$ standard error over five independent repetitions. Lower is better. *DGP* denotes the true data-generating conditional law and is therefore an oracle reference, not a fitted method. *Latent engression* is the infeasible benchmark trained directly on latent event times. Among feasible methods, the best result in each censoring regime is shown in bold.

Censoring	Method	Censored ES	Latent ES
Administrative	DGP	0.6332 ± 0.0061	1.2699 ± 0.0183
Administrative	Latent engression	0.6424 ± 0.0055	1.2862 ± 0.0182
Administrative	Naive observed engression	0.6444 ± 0.0053	1.3941 ± 0.0180
Administrative	Censored engression	$\mathbf{0.6401 \pm 0.0053}$	$\mathbf{1.3237 \pm 0.0193}$
Administrative	Independent Weibull	0.7850 ± 0.0033	1.4765 ± 0.0165
Administrative	MLP independent Weibull	0.7029 ± 0.0044	1.3772 ± 0.0175
Administrative	Clayton–Weibull	0.7586 ± 0.0027	1.4515 ± 0.0175
Administrative	MLP Clayton–Weibull	0.6745 ± 0.0055	1.3388 ± 0.0190
Administrative	Gumbel–Weibull	0.7567 ± 0.0025	1.4509 ± 0.0172
Administrative	Joint grid likelihood	0.7143 ± 0.0069	1.4761 ± 0.0156
Uniform	DGP	0.4840 ± 0.0067	1.2468 ± 0.0168
Uniform	Latent engression	0.4907 ± 0.0061	1.2601 ± 0.0169
Uniform	Naive observed engression	0.5524 ± 0.0125	1.4440 ± 0.0155
Uniform	Censored engression	$\mathbf{0.4904 \pm 0.0073}$	$\mathbf{1.2887 \pm 0.0207}$
Uniform	Independent Weibull	0.5975 ± 0.0071	1.4703 ± 0.0198
Uniform	MLP independent Weibull	0.5333 ± 0.0054	1.3532 ± 0.0208
Uniform	Clayton–Weibull	0.5781 ± 0.0070	1.4426 ± 0.0200
Uniform	MLP Clayton–Weibull	0.5152 ± 0.0070	1.3233 ± 0.0207
Uniform	Gumbel–Weibull	0.5773 ± 0.0071	1.4409 ± 0.0194
Uniform	Joint grid likelihood	0.5245 ± 0.0091	1.3586 ± 0.0183
Conditional uniform	DGP	0.3811 ± 0.0053	1.2468 ± 0.0168
Conditional uniform	Latent engression	0.3868 ± 0.0049	1.2601 ± 0.0169
Conditional uniform	Naive observed engression	0.4553 ± 0.0104	1.5009 ± 0.0199
Conditional uniform	Censored engression	$\mathbf{0.3871 \pm 0.0055}$	$\mathbf{1.3088 \pm 0.0236}$
Conditional uniform	Independent Weibull	0.4774 ± 0.0056	1.4879 ± 0.0214
Conditional uniform	MLP independent Weibull	0.4178 ± 0.0048	1.3698 ± 0.0217
Conditional uniform	Clayton–Weibull	0.4633 ± 0.0053	1.4640 ± 0.0206
Conditional uniform	MLP Clayton–Weibull	0.4051 ± 0.0057	1.3399 ± 0.0198
Conditional uniform	Gumbel–Weibull	0.4620 ± 0.0053	1.4628 ± 0.0209
Conditional uniform	Joint grid likelihood	0.4124 ± 0.0066	1.3582 ± 0.0219

Table 9: Synthetic multivariate survival experiment under a covariate-dependent mixture log-normal latent DGP with $k = 3$ event times and one shared censoring time. Entries are mean $\pm$ standard error over five independent repetitions. Lower is better. *DGP* denotes the true data-generating conditional law and is therefore an oracle reference, not a fitted method. *Latent engression* is the infeasible benchmark trained directly on latent event times. Among feasible methods, the best result in each censoring regime is shown in bold.

Censoring	Method	Censored ES	Latent ES
Administrative	DGP	0.9189 ± 0.0055	3.2238 ± 0.0389
Administrative	Latent engression	0.9457 ± 0.0044	3.2923 ± 0.0416
Administrative	Naive observed engression	0.9486 ± 0.0050	4.5247 ± 0.0667
Administrative	Censored engression	$\mathbf{0.9434 \pm 0.0039}$	3.7475 ± 0.0427
Administrative	Independent Weibull	1.1043 ± 0.0067	3.8241 ± 0.0559
Administrative	MLP independent Weibull	1.0809 ± 0.0056	4.0990 ± 0.0955
Administrative	Clayton–Weibull	0.9905 ± 0.0045	3.6364 ± 0.0520
Administrative	MLP Clayton–Weibull	0.9817 ± 0.0019	$\mathbf{3.5207 \pm 0.0361}$
Uniform	DGP	0.7451 ± 0.0073	3.1883 ± 0.0268
Uniform	Latent engression	0.7651 ± 0.0059	3.2389 ± 0.0296
Uniform	Naive observed engression	0.9003 ± 0.0089	4.4224 ± 0.0810
Uniform	Censored engression	$\mathbf{0.7584 \pm 0.0071}$	3.5878 ± 0.0608
Uniform	Independent Weibull	0.8922 ± 0.0086	3.6567 ± 0.0421
Uniform	MLP independent Weibull	0.8724 ± 0.0085	3.8675 ± 0.1367
Uniform	Clayton–Weibull	0.8059 ± 0.0085	3.4878 ± 0.0338
Uniform	MLP Clayton–Weibull	0.8014 ± 0.0078	$\mathbf{3.4073 \pm 0.0359}$
Conditional uniform	DGP	0.5965 ± 0.0073	3.1883 ± 0.0268
Conditional uniform	Latent engression	0.6137 ± 0.0060	3.2389 ± 0.0296
Conditional uniform	Naive observed engression	0.7489 ± 0.0113	4.5448 ± 0.0852
Conditional uniform	Censored engression	$\mathbf{0.6089 \pm 0.0061}$	3.6355 ± 0.0626
Conditional uniform	Independent Weibull	0.7155 ± 0.0087	3.6283 ± 0.0399
Conditional uniform	MLP independent Weibull	0.7006 ± 0.0086	3.7896 ± 0.0984
Conditional uniform	Clayton–Weibull	0.6463 ± 0.0084	3.4849 ± 0.0396
Conditional uniform	MLP Clayton–Weibull	0.6416 ± 0.0080	$\mathbf{3.4174 \pm 0.0448}$

Table 10: Synthetic multivariate survival experiment under a covariate-dependent mixture log-normal latent DGP with $k = 5$ event times and one shared censoring time. Entries are mean $\pm$ standard error over five independent repetitions. Lower is better. *DGP* denotes the true data-generating conditional law and is therefore an oracle reference, not a fitted method. *Latent engression* is the infeasible benchmark trained directly on latent event times. Among feasible methods, the best result in each censoring regime is shown in bold.

Censoring	Method	Censored ES	Latent ES
Administrative	DGP	1.2108 ± 0.0129	5.3617 ± 0.0205
Administrative	Latent engression	1.3204 ± 0.0272	5.5121 ± 0.0363
Administrative	Naive observed engression	1.2445 ± 0.0129	7.9679 ± 0.0529
Administrative	Censored engression	$\mathbf{1.2258 \pm 0.0121}$	7.1165 ± 0.0267
Administrative	Independent Weibull	1.4678 ± 0.0119	6.1825 ± 0.0446
Administrative	MLP independent Weibull	1.4394 ± 0.0112	6.6426 ± 0.1223
Administrative	Clayton–Weibull	1.3084 ± 0.0122	5.9206 ± 0.0317
Administrative	MLP Clayton–Weibull	1.3140 ± 0.0167	$\mathbf{5.8265 \pm 0.0393}$
Uniform	DGP	0.9536 ± 0.0100	5.1394 ± 0.0078
Uniform	Latent engression	1.0477 ± 0.0204	5.3048 ± 0.0275
Uniform	Naive observed engression	1.1433 ± 0.0070	7.6696 ± 0.0984
Uniform	Censored engression	$\mathbf{0.9692 \pm 0.0110}$	6.6095 ± 0.0641
Uniform	Independent Weibull	1.1529 ± 0.0132	5.9272 ± 0.0301
Uniform	MLP independent Weibull	1.1326 ± 0.0132	6.1641 ± 0.0837
Uniform	Clayton–Weibull	1.0356 ± 0.0117	5.6937 ± 0.0365
Uniform	MLP Clayton–Weibull	1.0441 ± 0.0120	$\mathbf{5.5976 \pm 0.0432}$
Conditional uniform	DGP	0.7638 ± 0.0105	5.1394 ± 0.0078
Conditional uniform	Latent engression	0.8536 ± 0.0192	5.3048 ± 0.0275
Conditional uniform	Naive observed engression	0.9440 ± 0.0064	7.8476 ± 0.0988
Conditional uniform	Censored engression	$\mathbf{0.7754 \pm 0.0117}$	6.7030 ± 0.0690
Conditional uniform	Independent Weibull	0.9230 ± 0.0137	5.9580 ± 0.0362
Conditional uniform	MLP independent Weibull	0.9022 ± 0.0137	6.2551 ± 0.0646
Conditional uniform	Clayton–Weibull	0.8308 ± 0.0124	5.7525 ± 0.0428
Conditional uniform	MLP Clayton–Weibull	0.8392 ± 0.0148	$\mathbf{5.6475 \pm 0.0527}$

Table 11: Synthetic multivariate survival experiment under a covariate-dependent mixture log-normal latent DGP with $k = 10$ event times and one shared censoring time. Entries are mean $\pm$ standard error over five independent repetitions. Lower is better. *DGP* denotes the true data-generating conditional law and is therefore an oracle reference, not a fitted method. *Latent engression* is the infeasible benchmark trained directly on latent event times. Among feasible methods, the best result in each censoring regime is shown in bold.

Censoring	Method	Censored ES	Latent ES
Administrative	DGP	1.6530 ± 0.0092	9.6954 ± 0.0843
Administrative	Latent engression	2.0170 ± 0.0403	10.2168 ± 0.0699
Administrative	Naive observed engression	1.7181 ± 0.0098	17.3182 ± 0.2313
Administrative	Censored engression	$\mathbf{1.6884 \pm 0.0107}$	15.7127 ± 0.1847
Administrative	Independent Weibull	2.0100 ± 0.0088	14.9379 ± 0.6241
Administrative	MLP independent Weibull	1.9585 ± 0.0118	13.3928 ± 0.2939
Administrative	Clayton–Weibull	1.8086 ± 0.0074	14.3108 ± 0.5964
Administrative	MLP Clayton–Weibull	1.8793 ± 0.0186	$\mathbf{13.1867 \pm 0.2063}$
Uniform	DGP	1.3131 ± 0.0055	9.4613 ± 0.0914
Uniform	Latent engression	1.6082 ± 0.0190	10.0111 ± 0.1553
Uniform	Naive observed engression	1.6402 ± 0.0228	16.5122 ± 0.3144
Uniform	Censored engression	$\mathbf{1.3382 \pm 0.0052}$	14.2576 ± 0.3110
Uniform	Independent Weibull	1.5901 ± 0.0086	13.9407 ± 0.4351
Uniform	MLP independent Weibull	1.5491 ± 0.0066	$\mathbf{12.4447 \pm 0.1425}$
Uniform	Clayton–Weibull	1.4397 ± 0.0073	13.3828 ± 0.4909
Uniform	MLP Clayton–Weibull	1.4783 ± 0.0200	12.7285 ± 0.5428
Conditional uniform	DGP	1.0562 ± 0.0067	9.4613 ± 0.0914
Conditional uniform	Latent engression	1.3390 ± 0.0175	10.0111 ± 0.1553
Conditional uniform	Naive observed engression	1.3623 ± 0.0180	16.8903 ± 0.3222
Conditional uniform	Censored engression	$\mathbf{1.0815 \pm 0.0055}$	14.6636 ± 0.2549
Conditional uniform	Independent Weibull	1.2751 ± 0.0089	13.7644 ± 0.4093
Conditional uniform	MLP independent Weibull	1.2406 ± 0.0083	12.7779 ± 0.1484
Conditional uniform	Clayton–Weibull	1.1574 ± 0.0079	13.2792 ± 0.4230
Conditional uniform	MLP Clayton–Weibull	1.1837 ± 0.0091	$\mathbf{12.5259 \pm 0.4063}$

Table 12: Synthetic multivariate survival experiment under a regular log-normal latent DGP with $k = 2$ event times and one shared censoring time. Entries are mean $\pm$ standard error over five independent repetitions. Lower is better. *DGP* denotes the true data-generating conditional law and is therefore an oracle reference, not a fitted method. *Latent engression* is the infeasible benchmark trained directly on latent event times. Among feasible methods, the best result in each censoring regime is shown in bold.

Censoring	Method	Censored ES	Latent ES
Administrative	DGP	0.4288 ± 0.0040	0.7176 ± 0.0069
Administrative	Latent engression	0.4305 ± 0.0048	0.7196 ± 0.0063
Administrative	Naive observed engression	0.4416 ± 0.0035	0.8642 ± 0.0135
Administrative	Censored engression	$\mathbf{0.4308 \pm 0.0041}$	0.7892 ± 0.0123
Administrative	Independent Weibull	0.4365 ± 0.0039	0.7268 ± 0.0071
Administrative	MLP independent Weibull	0.4380 ± 0.0049	0.7403 ± 0.0081
Administrative	Clayton–Weibull	0.4329 ± 0.0038	$\mathbf{0.7225 \pm 0.0069}$
Administrative	MLP Clayton–Weibull	0.4360 ± 0.0051	0.7299 ± 0.0091
Administrative	Gumbel–Weibull	0.4340 ± 0.0040	0.7236 ± 0.0071
Administrative	Joint grid likelihood	0.4506 ± 0.0037	0.9149 ± 0.0167
Uniform	DGP	0.3221 ± 0.0056	0.7175 ± 0.0117
Uniform	Latent engression	0.3238 ± 0.0060	0.7259 ± 0.0119
Uniform	Naive observed engression	0.3975 ± 0.0088	0.9709 ± 0.0211
Uniform	Censored engression	0.3257 ± 0.0055	0.7886 ± 0.0249
Uniform	Independent Weibull	0.3275 ± 0.0054	0.7259 ± 0.0127
Uniform	MLP independent Weibull	0.3290 ± 0.0063	0.7421 ± 0.0138
Uniform	Clayton–Weibull	$\mathbf{0.3251 \pm 0.0053}$	0.7229 ± 0.0128
Uniform	MLP Clayton–Weibull	0.3280 ± 0.0064	0.7395 ± 0.0126
Uniform	Gumbel–Weibull	0.3257 ± 0.0055	$\mathbf{0.7225 \pm 0.0128}$
Uniform	Joint grid likelihood	0.3407 ± 0.0057	0.8191 ± 0.0204
Conditional uniform	DGP	0.2672 ± 0.0055	0.7175 ± 0.0117
Conditional uniform	Latent engression	0.2687 ± 0.0057	0.7259 ± 0.0119
Conditional uniform	Naive observed engression	0.3557 ± 0.0098	1.0458 ± 0.0231
Conditional uniform	Censored engression	$\mathbf{0.2694 \pm 0.0057}$	0.7927 ± 0.0198
Conditional uniform	Independent Weibull	0.2717 ± 0.0051	0.7266 ± 0.0128
Conditional uniform	MLP independent Weibull	0.2729 ± 0.0061	0.7504 ± 0.0147
Conditional uniform	Clayton–Weibull	0.2699 ± 0.0053	0.7235 ± 0.0125
Conditional uniform	MLP Clayton–Weibull	0.2729 ± 0.0064	0.7478 ± 0.0132
Conditional uniform	Gumbel–Weibull	0.2703 ± 0.0053	$\mathbf{0.7233 \pm 0.0129}$
Conditional uniform	Joint grid likelihood	0.2798 ± 0.0057	0.8203 ± 0.0199

C.7 ADDITIONAL AKI USE-CASE SETUP AND RESULTS

This subsection describes the ICU AKI use case used in Section 5.3. The goal is methodological: we use a realistic, irregularly sampled ICU prediction problem to test whether a censoring-aware distributional objective improves observed-data predictive distributions. We therefore report processed cohort sizes and endpoint rates, but do not include a demographics table.

Table 13: KDIGO Acute Kidney Injury (AKI) stage definitions for serum creatinine (sCr) and urine output (UO) used in this work.

Stage	sCr criteria	UO criteria
Stage 1	sCr $\geq 1.5 - 1.9\times$ baseline, or increase in sCr ≥ 0.3 mg/dL	UO ≤ 0.5 mL/kg/h for 6–12 h
Stage 2	sCr $\geq 2.0 - 2.9\times$ baseline	UO ≤ 0.5 mL/kg/h for ≥ 12 h
Stage 3	sCr $\geq 3.0\times$ baseline, or sCr ≥ 4.0 mg/dL	UO ≤ 0.3 mL/kg/h for ≥ 24 h

Cohort construction and prediction rows. We use MIMIC-IV ICU stays and construct one prediction row per eligible ICU stay and timestamp. Prediction rows are kept only when they occur within the first 14 days after ICU admission and have positive remaining follow-up time before ICU exit (discharge alive or in-ICU death). The two event times are the remaining times from the prediction timestamp to KDIGO (KDIGO, 2012) stage-2 AKI according to the creatinine criterion and according to the urine-output criterion, see Table 13. The shared censoring time C is the remaining time to ICU exit. For endpoint $j \in \{1, 2\}$,

$$Y_j = \min(T_j, C), \qquad \Delta_j = \mathbb{1}\{T_j \leq C\},$$

where T_j is the latent time to the corresponding AKI definition. Times are transformed during training as

$$t \mapsto \frac{\log(1+t)}{\log(1+q_{0.95})},$$

where $q_{0.95}$ is the 95th percentile of the training ICU-exit-time distribution. All reported localized energy scores and fixed-horizon metrics are computed after transforming sampled times back to hours.

The split is chronological at the ICU-stay level. Stays from anchor-year groups 2008–2016, plus one-half of the 2017–2019 stays, are used for training; the remaining 2017–2019 stays are used for validation; 2020–2022 stays are held out for testing. This yields $4{,}831{,}810$ training rows, $575{,}500$ validation rows, and $835{,}661$ test rows. The training row-level event rates are 18.3% for creatinine-defined AKI and 39.7% for urine-output-defined AKI; however, note that this is an artificially high value because stays in which the event has already occurred are also counted as events. We do so because urine-defined AKI may already have occurred; however, creatinine-defined AKI may still occur, and we still want to model it. Additionally, one can exit the event at the next time point. During training, the loss function is set to zero for events that have already occurred, since this can be inferred from the observation at that moment. The median remaining ICU-exit time is 54.7 hours.

Covariates and preprocessing. Feature construction follows the ICU preprocessing pipeline of Jonkers et al. (2026), adapted to the covariates used in this study. Each prediction row has 22 standardized covariates, including gender, age at admission, weight, chronic renal failure status, baseline serum creatinine, elapsed ICU time, and 72-hour summary features for serum creatinine and urine output: count, mean, minimum, maximum, latest value, and slope. We additionally include missingness and slope-availability indicators for the two time-varying measurements. Imputation and scaling are fit on the training split only: missing counts are set to zero, missing measurement summaries are filled with training medians, missing slopes are set to zero, and all model features are standardized using training means and standard deviations.

Training objectives and baselines. The censored engression model is a noise-conditioned MLP generator with noise dimension 500, hidden width 500, depth 2, and a softplus output transform. It is trained with AdamW, learning rate $2 \cdot 10^{-4}$, weight decay 10^{-5}, batch size 1024, gradient clipping at 1, and early stopping on the validation localized censored energy score. Each training update uses 16 generated samples. A stay-aware minibatch sampler prevents two rows from the same ICU stay from appearing in the same minibatch. The naive engression baseline uses the same generator class but replaces the localized censored objective by an ordinary energy-score objective that treats the observed censored vector Y as if it were the true event-time vector.

We compare against five likelihood-based baselines trained on the same processed tensors: independent Weibull, conditional MLP independent Weibull, Clayton copula Weibull, conditional MLP Clayton copula Weibull, and a joint discrete-grid likelihood model with 32 bins. The conditional Weibull and conditional Clayton baselines use hidden width 500 and depth 2; the joint grid baseline is included since this AKI task is bivariate.

Evaluation. Since the ICU-exit censoring time is observed for every prediction row, we evaluate using localized scores. For the bivariate energy score, generated samples are first localized by

$$\psi_C^\flat(z_1, z_2) = (\min(z_1, C), \min(z_2, C)),$$

and are then compared with the observed censored vector Y. We report the joint localized energy score for rows with at least one endpoint still at risk, and the marginal localized energy scores after excluding rows in which the corresponding endpoint is already present at prediction time. The test set contributes $802{,}657$ rows to the joint score, $725{,}596$ rows to the creatinine marginal score, and $699{,}383$ rows to the urine-output marginal score.

We also compute localized Brier scores at fixed horizons for the composite endpoint "either AKI". For a horizon τ, the censored predicted risk is set to one when $C \le \tau$, and otherwise uses the latent sampled risk, e.g.

$$F_{C,\tau}^\flat(x) = \begin{cases} 1, & C \le \tau, \\ \Pr_F(T_1 \le \tau \text{ or } T_2 \le \tau \mid X = x), & C > \tau. \end{cases}$$

The observed censored binary outcome is $\mathbb{1}\{Y_1 \le \tau \text{ or } Y_2 \le \tau\}$. Top-$5\%$ and top-10% positive predictive value and sensitivity are reported only as secondary ranking summaries on the horizon-specific evaluable subset. They are not valid scores and should not be interpreted as estimates of deployment utility.

Fixed-horizon results. Table 14 reports the composite endpoint at 6, 12, 24, and 48 hours. Censored engression obtains the best localized Brier score at all four horizons. The gap is smallest at 6 hours and becomes more pronounced from 12 to 48 hours, where censoring by ICU exit is more consequential. The secondary enrichment metrics are directionally consistent with the score results: censored engression usually selects a higher-risk top tail than the naive generator and the likelihood baselines. These summaries are useful for clinical interpretation, but the localized Brier score is the primary fixed-horizon criterion.

Distributional results. Table 15 compares methods that can generate full bivariate event-time samples. Censored engression has the lowest joint localized energy score and the lowest marginal localized energy scores, which coincide with localized CRPS in the univariate $\beta = 1$ case, for both AKI definitions. The joint discrete-grid likelihood baseline is the closest competitor, which suggests that flexible joint modeling is important for this task; however, it remains slightly worse than censored engression and is tied to the low-dimensional bivariate setting. Conditional Weibull and conditional Clayton baselines substantially improve upon their non-conditional counterparts, confirming the importance of covariate-dependent event-time distributions, but they do not match the sample-based censored engression objective. Overall, the AKI use case supports the same methodological conclusion as the synthetic experiments: treating censored observations as fully observed event times can distort the learned distribution, whereas localizing the score at the observed censoring time yields better predictive distributions based on the observed data.

Interpretational caveats. The row-level event rates above are prediction-row rates, not patient-level incidence estimates, because the same ICU stay can contribute multiple prediction times. In addition,

Table 14: Secondary utility metrics for the composite endpoint (either AKI). Localized Brier is a score, so lower is better; PPV and sensitivity are utility metrics, so higher is better. These metrics are descriptive only. Bold entries indicate the best method within each horizon and metric.

Horizon	Method	Loc. Brier ↓	PPV@5% ↑	Sens.@5% ↑	PPV@10% ↑	Sens.@10% ↑
6h	Naive	0.023	0.485	0.528	0.338	0.735
6h	Censored	**0.022**	**0.501**	**0.545**	**0.348**	**0.757**
6h	Joint grid	0.028	0.465	0.505	0.341	0.742
6h	Indep. Weibull	0.223	0.177	0.193	0.157	0.342
6h	Clayton Weibull	0.206	0.179	0.194	0.159	0.345
6h	Cond. indep. Weibull	0.041	0.403	0.439	0.313	0.680
6h	Cond. Clayton Weibull	0.047	0.366	0.398	0.290	0.632
12h	Naive	0.045	0.727	0.316	0.615	0.535
12h	Censored	**0.040**	**0.782**	**0.340**	**0.664**	**0.578**
12h	Joint grid	0.043	0.728	0.317	0.622	0.542
12h	Indep. Weibull	0.216	0.367	0.160	0.327	0.284
12h	Clayton Weibull	0.201	0.369	0.160	0.330	0.287
12h	Cond. indep. Weibull	0.051	0.655	0.285	0.560	0.488
12h	Cond. Clayton Weibull	0.054	0.633	0.275	0.547	0.476
24h	Naive	0.081	0.759	0.167	0.696	0.307
24h	Censored	**0.071**	**0.840**	**0.185**	**0.751**	**0.331**
24h	Joint grid	0.074	0.805	0.177	0.724	0.319
24h	Indep. Weibull	0.190	0.514	0.113	0.470	0.207
24h	Clayton Weibull	0.179	0.515	0.114	0.473	0.208
24h	Cond. indep. Weibull	0.077	0.762	0.168	0.690	0.304
24h	Cond. Clayton Weibull	0.078	0.757	0.167	0.689	0.304
48h	Naive	0.095	0.769	0.094	0.763	0.186
48h	Censored	**0.080**	**0.929**	**0.113**	**0.882**	**0.215**
48h	Joint grid	0.081	0.911	0.111	0.862	0.211
48h	Indep. Weibull	0.138	0.676	0.083	0.648	0.158
48h	Clayton Weibull	0.132	0.678	0.083	0.649	0.158
48h	Cond. indep. Weibull	0.083	0.878	0.107	0.831	0.203
48h	Cond. Clayton Weibull	0.083	0.886	0.108	0.840	0.205

Table 15: Additional AKI use-case benchmark results using localized energy scores. Lower is better. These scores evaluate overall fit of the predicted event-time distributions.

Method	Joint localized ES	Creatinine localized ES	Urine localized ES
Naive observed engression	43.442	23.591	32.981
Censored engression	**30.933**	**15.439**	**23.769**
Independent Weibull	45.948	25.915	34.154
Conditional MLP independent Weibull	35.728	19.534	26.061
Clayton copula + Weibull	45.749	25.915	34.154
Conditional MLP Clayton copula + Weibull	34.006	18.863	24.606
Joint discrete-grid likelihood	31.675	16.210	24.057

ICU discharge is an observed exit process rather than a randomized censoring mechanism. The reported localized scores, therefore, evaluate the predictive distribution under the stated observed-data censoring convention; they should not be interpreted as causal estimates of latent AKI incidence after discharge. The use case is intended to demonstrate the proposed scoring and training methodology, not to define a final clinical deployment model.