Task-Dependent Feature Representation in Long-Horizon Survival Forecasting: A 47-Year Manga Cohort Study

Ryoma Yasunaga
Georgia Institute of Technology
Atlanta, Georgia, USA
ryasunaga3@gatech.edu

Abstract

Long-running serial publication formats (weekly manga among them) generate decades of cancellation outcomes that a reader-facing forecasting tool, and secondarily an editorial-planning aid, must predict from past cohorts only. We evaluate survival models on a 47-year, four-magazine, 99,190-issue corpus under a prospective rolling-origin protocol with five landmarks, comparing six architectures (LSTM, Transformer, S3-Hybrid, RSF, DeepSurv, DeepHit) against seven feature families on the same training pool. The reader-facing deployment metric is lead-time AUROC at horizons $k \in \{1, \ldots, 26\}$; we contrast it with the conventional within-issue ranking metric, pool concordance C_{pool}. The two tell opposite stories about what features matter. Lead-time AUROC is led by LSTM/Transformer with hidden-Markov-trajectory or rule-and-placement features (H_{full}, HRP) at every horizon, exceeding the pool-concordance champion by +0.065 to +0.074 in mean AUROC across five seeds under matched bootstrap, and the strongest static survival configuration by +0.106 to +0.179 in 130/130 cells; pool concordance, by contrast, is solved near-ceiling by a vanilla LSTM on raw placement trajectories alone ($C_{\text{pool}} = 0.791$, 5-landmark mean), and engineered features in fact *degrade* this metric. The dissociation is *feature-class invariant*: training LSTM with lag-shifted, exponentially weighted, or Bayesian change-point representations reproduces the inversion across 14 feature cells (Spearman $\rho \in [-0.79, -0.52]$ over five landmarks, top-3 cells *disjoint* at every landmark), confirmed by 5-seed replication ($\rho < 0$ in 25/25 seed-landmark configurations). We formalise this theoretically; the implication for reader-facing deployment is that the lead-time-AUROC-optimal configurations are the appropriate choice.

CCS Concepts

• **Computing methodologies** → **Supervised learning by classification**; *Neural networks*; • **Theory of computation** → **Survival analysis**; • **Information systems** → Decision support systems.

Keywords

survival analysis, prospective evaluation, hidden Markov models, feature representation, time series, manga

ACM Reference Format:
Ryoma Yasunaga. 2026. Task-Dependent Feature Representation in Long-Horizon Survival Forecasting: A 47-Year Manga Cohort Study. In *Proceedings of The 12th International Workshop on Mining and Learning from Time Series, co-located with KDD 2026 (MILETS '26).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

MILETS '26, Jeju, Republic of Korea
2026. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

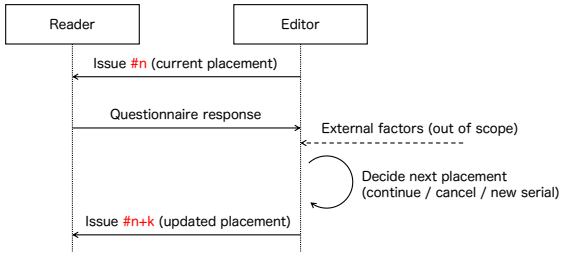

Figure 1: Weekly editor–reader feedback loop. A reader response to issue #n is reflected only at issue #n+k; out-of-scope factors (e.g., tankōbon sales) are shown dashed. This k-issue response-to-reflection lag is the forecast horizon we evaluate throughout.

1 Introduction

In Japan, weekly shōnen manga magazines are the upstream origin of major media-mix franchises spanning anime, film, and merchandise [31], and the continuation or cancellation of an individual serial in such a magazine is largely governed by reader questionnaire responses [19]. The signal is most decisive in the early pre-tankōbon weeks, when the questionnaire is the only channel through which readers can influence outcomes. Yet readers face a structural resource-allocation problem: each questionnaire admits only a small fixed slate of supported titles and the right to respond is tied to purchasing the issue. Deciding *when* to spend a vote on a niche favourite, rather than on a title whose continuation is nearly assured, is therefore a non-trivial decision under cancellation uncertainty. A forecast issued ahead of this decision window is valuable as a *reader-facing* decision aid before it is valuable to any editorial process.

These ecosystems run on a continuous reader–editor feedback loop (Fig. 1): a reader's response to issue #n is reflected only at issue #n+k, with the same loop running in parallel for the $k-1$ intervening issues. A forecast with k issues of lead time therefore needs to be evaluated at *multiple horizons* rather than one step ahead, in its primary role as a reader-facing decision aid (and secondarily as an editorial-planning input). Public archives such as the Media Arts Database [24] now expose a long-horizon, heavy-tailed survival problem at the work level; we adopt a prospective rolling-origin protocol that trains on past cohorts and tests on future ones.

The deployment metric that matches this reader-facing use case is *lead-time AUROC* at horizon k (AUROC_k), the detection of works that will be cancelled within k issues. The conventional alternative, issue-level *pool concordance* (C_{pool}), ranks co-appearing works

within a single issue and corresponds to an editorial use case (within-issue placement and cancellation decisions). Read as k-step forecasting and one-step ranking respectively, one might expect them to reward similar feature representations. Our central empirical finding is that they do not. Under matched bootstrap on a 99,190-issue, 47-year corpus across five prospective landmarks, the lead-time top tier is occupied by detail-preserving regime-encoded features, while the cell that wins under pool concordance (a vanilla LSTM on raw placement trajectories alone) is uniformly *absent* from it. Adding engineered features to the raw LSTM cell in fact *degrades* pool concordance, while removing them *lowers* lead-time AUROC.

We unify this as a claim about the *sufficient statistics* of the two tasks. Under a hidden-Markov regime-switching DGP, the k-step cumulative event probability factorises into a current-regime posterior and a k-step trajectory kernel (§4); pool concordance is dominated by the first factor under approximate stochastic ordering, while lead-time AUROC at non-trivial k requires both. The inversion is theoretically expected, and the empirical core of the paper documents it across 18 sequence cells, four feature classes, and 5-seed replication (§5–§7).

Contributions. **(C1) Theoretical characterisation.** We prove that, under a HMM regime-switching DGP, lead-time AUROC at horizon k admits a bilinear sufficient statistic decomposition $F_{\tau,k}(H_\tau) = 1 - \langle \mathbf{h}_\tau^{\text{snap}}, \mathbf{h}_\tau^{\text{traj}}(k) \rangle$, while pool concordance is maximised by a function of the current regime alone whenever conditional survival is stochastically ordered. The mismatch is therefore not an artefact of the corpus but a property of the metric pair. **(C2) Prospective evaluation protocol with apples-to-apples feature–architecture grid.** We evaluate at five landmarks $\tau \in \{1990, \ldots, 2010\}$ at 5-year spacing, on four magazines totalling 99,190 issues over 47 years, with six architectures crossed against seven feature families. Within-issue paired bootstrap with shared resamples across configurations enables matched-sample inference at every cell. **(C3) Inversion universality and replication.** The pool-vs-lead-time inversion reproduces uniformly across four mechanistically distinct feature classes (HMM, lag-k, EWMA-α, BOCPD) and uniformly across five replication seeds; we report this with full per-seed AUROC dispersion and a counterfactual-deletion sanity check that qualitatively but not strictly confirms the trajectory-factor contribution at the deep-model level.

2 Related Work

Survival of cultural products. Bradlow and Fader [6] introduced Bayesian lifetime modelling of Billboard chart trajectories; De Vany and Walls [10] documented heavy-tailed box-office survival in the movie industry; Bhattacharjee et al. [5] applied Cox proportional-hazards models to album chart longevity; Candia et al. [7] characterised a universal decay of collective attention across cultural domains. These works characterise cultural decline on completed runs rather than the weekly editorial cancellation hazard that drives serial publication.

Hidden Markov models and regime-switching time series. HMMs [11, 12, 23, 28] are the canonical tool for inferring latent state from observable signals over time. Bayesian online change-point detection [1] provides a complementary regime-shift representation. We use HMMs as the primary representation and lag-k/EWMA/BOCPD as comparators to test whether the central empirical finding (pool-vs-lead-time inversion) is specific to HMM structure or holds across regime-aware feature classes more generally.

Discrimination metrics for survival models. The pool concordance index extends the classical concordance of Harrell et al. [13] by restricting comparisons to within-issue pairs; its time-dependent counterparts [2] and the binary-classification view of survival via lead-time AUROC [14] are the two metrics we contrast. The contrast itself, that different survival evaluation metrics admit different sufficient statistics and that a single best score function generally does not exist, is the object of our theoretical analysis (§6).

Prospective rolling-origin evaluation. Tashman [32], Bergmeir and Benítez [3], and Bergmeir et al. [4] argued that random and simple temporal splits give systematically optimistic estimates under future deployment and that rolling-origin evaluation aligned with the deployment scenario is more informative; the view is reinforced in textbook treatments of forecasting [15]. We adopt a five-landmark prospective rolling-origin protocol that conditions every reported metric on the deployment-style train/test separation.

3 Data and Methods

3.1 Dataset

The corpus comprises four major Japanese weekly manga magazines spanning 47 years (1969–2016), assembled from the bibliographic table-of-contents records of the Media Arts Database [24], an open archive maintained by the National Center for Art Research and made freely available for reuse under the database's own terms of use, which request source and version attribution. Each row is an *issue record*: one appearance of a work in a single weekly issue of one magazine. The full corpus contains 99,190 issue records that resolve to a few thousand distinct works. We use two within-issue position statistics throughout. The *continuous within-issue placement* $\text{placement}_t \in (0, 1]$ is the chapter's leading-page number divided by the issue's total page count; the integer *within-issue rank* $\text{rank\_in\_issue}_t \in \{1, \ldots, n_t\}$ is the ordinal position induced by placement_t within issue t. Each record additionally carries the page count of the chapter and metadata identifying the magazine and the year.

We define cancellation as the terminal event of a serial: the issue at which the work last appears. Right-censoring is restricted to works whose final appearance falls within the last calendar year of our window (2016), preventing leakage of unobserved cancellations.

3.2 Feature families

We organise the covariate space into seven families that span the spectrum from raw observation through editor-side metadata, rule-based heuristics, and three increasingly detail-preserving regime-encoding representations, letting us trace how the two metrics respond to a controlled increase in trajectory information. We use τ for the deployment landmark, t for the current issue index, and

placement_ma4 for the four-issue moving average of the continuous placement placement$_t$ (*not* of the ordinal rank_in_issue$_t$).

MIN (3 features). We feed sequence models the per-issue triple (rank_in_issue$_t$, pages$_t$, magazine) at every observed $t \le \tau$, as the unprocessed trajectory. The triple is the smallest representation an editor would have on hand at deployment, and feeding it without further preprocessing isolates the contribution of the sequence cell's internal representation from any handcrafted aggregation. Static survival models do not accept this family because they cannot consume variable-length per-issue trajectories without an external summary.

BASE (5 features). Five work-level covariates that an editor inspects at first appearance: magazine identifier, start year, initial placement (continuous placement$_t$ at first appearance, t=1), mean page count across observed issues, and a five-bucket era indicator (1970s, 1980s, 1990s, 2000s, 2010s) that absorbs slow-changing genre and serialisation norms. All five are time-invariant by construction, so BASE feeds both static and sequence cells without modification.

R (5 features). Five rule-derived alert and streak features that mimic the heuristic checks editors use at the desk. An alert fires at issue t when placement_ma4 exceeds a threshold θ for N_{consec} consecutive issues; from the resulting alert sequence we derive (i) a binary flag for whether any alert has fired by t, (ii) the issue index of the first alert (-1 if none), (iii) the cumulative alert count, (iv) the current consecutive-exceedance streak length at t, and (v) the maximum streak length observed by t. The thresholds $(\theta, N_{\text{consec}}) = (0.5, 2)$ are oracle-tuned on the prior training pool. R is the lowest-capacity engineered family in our grid and acts as a numerically interpretable comparator for the HMM families.

H_{snap} (5 features). Five hidden-Markov snapshot statistics evaluated at the most recent observed issue $t = \tau$: (i) the regime-mixture cancellation probability $\sum_{l=1}^{K} P(s_t = l \mid H_t) \cdot \tilde{h}(l)$, (ii) the ARGMAX most-likely state $\arg\max_l P(s_t = l \mid H_t)$, (iii) the four-issue moving average of statistic (i) over $t-3 \le t' \le t$, (iv) the state-posterior entropy $-\sum_l P(s_t = l \mid H_t) \log P(s_t = l \mid H_t)$, and (v) the count of MAP state changes $|\{t' \le t : \arg\max_l P(s_{t'} \mid H_{t'}) \neq \arg\max_l P(s_{t'-1} \mid H_{t'-1})\}|$. Statistics (i)–(iv) probe the snapshot factor $\mathbf{h}_\tau^{\text{snap}} = P(s_\tau \mid H_\tau)$ of Theorem 2; (v) measures regime instability, which is near-zero under strict regime persistence and grows when stochastic ordering of $\{S_t\}$ would fail.

H_{traj} (5 features). Five aggregates of the full posterior trajectory $\{P(s_t \mid H_t)\}_{t \le \tau}$: (i) the trajectory mean and (ii) trajectory max of the regime-mixture cancellation probability, (iii) the fraction of issues at which the MAP state is the highest hazard state ($l^* = \arg\max_l \tilde{h}(l)$, in our fits the rightmost of the four states), (iv) the issue index at which the work first entered that state (or τ+1 if never), and (v) the longest run of consecutive issues spent in the highest-hazard state. These five features summarise the trajectory factor $\mathbf{h}_\tau^{\text{traj}}(k) = \tilde{A}^k \mathbf{1}$ of Theorem 2 in a horizon-agnostic form: they capture how the regime path has accumulated risk up to τ rather than how the posterior was distributed at the instant τ.

H_{full} (10 features). $H_{\text{full}} = H_{\text{snap}} \cup H_{\text{traj}}$. By construction this is the empirical analogue of the bilinear factorisation $F_{\tau,k} = 1 - \langle \mathbf{h}_\tau^{\text{snap}}, \mathbf{h}_\tau^{\text{traj}}(k) \rangle$ of Theorem 2: a model given H_{full} has direct access to a low-dimensional summary of both factors and need only learn the inner-product weighting appropriate to the horizon. The empirical advantage of H_{full} at long k relative to H_{snap} or H_{traj} alone (§5.2) is the cell-level signature predicted by the theorem.

HRP.. The widest engineered family, constructed as HRP = BASE$\cup$ $H_{\text{full}} \cup$ R $\cup$ P, where P is a 7-element placement-trajectory pack of cumulative trajectory statistics computed over $1 \le t' \le \tau$: (i) placement_ma4 at $t = \tau$, (ii) its cumulative mean, (iii) cumulative max, and (iv) cumulative min over the trajectory; (v) the cumulative mean of the 4-issue rolling standard deviation placement_sd4; (vi) the fraction of issues at which placement_ma4 has exceeded 0.5; and (vii) a half-period trend statistic (second-half mean minus first-half mean of placement_ma4). HRP thus combines all editor metadata, all rule heuristics, both HMM factors, and a mechanistic non-HMM representation of placement dynamics in one feature vector, and is the highest-capacity static-model input in our grid.

Comparator regime-encoding classes. For the feature-class invariance analysis of §5.3 we additionally compute three non-HMM regime-encoding feature classes on the same protocol: (a) lag-shifted placement placement_ma4$_{t-k_{\text{lag}}}$ at $k_{\text{lag}} \in \{3, 6, 12\}$, three cells; (b) exponentially weighted moving averages of placement_ma4 with smoothing parameter $\alpha \in \{0.1, 0.3, 0.5\}$, three cells; and (c) a single Bayesian online change-point cell using the run-length posterior of Adams and MacKay [1] with a Gaussian observation model on placement_ma4 and a constant hazard prior of $1/\lambda_h = 30$ issues. These three classes are mechanistically distinct from the HMM (no latent state, no killed-transition algebra) and serve as alternative parametrisations of trajectory dynamics.

HMM hyperparameters. The HMM is fit by EM on bivariate placement_ma4 and placement_sd4 sequences of the training pool at each landmark, with K=4 latent states, Gaussian emissions with full covariance, a row-stochastic transition matrix without sparsity constraints, and initial distribution, transition matrix, and emission parameters jointly learned by EM. K=4 is selected by BIC elbow analysis on the τ=1990 training pool, with each of the four states carrying stationary mass $\ge 5\,\%$ under the fitted transition's stationary distribution. A state-count sweep ($K \in \{2, \ldots, 8\}$, §7.2) confirms that K=4 is not load-bearing. The HMM is fit once per landmark on the training pool only; the prospective test cohort is scored by the forward algorithm with frozen parameters, so no test-cohort information enters fitting.

3.3 Prospective rolling-origin protocol

For each landmark $\tau \in \{1990, 1995, 2000, 2005, 2010\}$ we partition works by their start year: all works that began strictly before τ form the training pool, and all works that began on or after τ form the prospective test cohort. Hyperparameters are pre-validated once on the τ=1990 training pool by five-fold cross-validation; the selected hyperparameters are frozen and applied at all five landmarks, eliminating per-landmark hyperparameter tuning that would contaminate prospective comparison.

All reported metrics use a paired bootstrap with B=500 resamples drawn at the issue level with a fixed seed, so that pairwise

differences across configurations within a landmark are evaluated under matched samples [3, 4, 32].

3.4 Metrics

Issue-level pool concordance. C_{pool} extends the classical concordance index of Harrell et al. [13] by restricting the ranking comparison to pairs of works that co-occur in the same issue of the same magazine, and pooling these comparisons across all issues of the test cohort. This adapts the discrimination metric to our deployment unit, namely the editorial decision context inside a single issue, and is closer in spirit to the time-dependent discrimination index of Antolini et al. [2] than to the unrestricted concordance.

Lead-time AUROC.. For each lead-time k we form the time-dependent binary classification task "does the work cancel within k issues from the snapshot?" over a rolling per-issue snapshot grid, and compute the AUROC of each model's predicted hazard against this label [14].

Antolini's C as a sanity check. For §7 we additionally compute Antolini's IPCW-weighted time-dependent C-index [2] aggregated to the work level. Pool C_{pool} and Antolini's C are only moderately correlated across the LSTM universality cells (median Spearman $\rho \approx 0.43$ across $(\tau, \text{horizon})$ pairs), but *both* disagree with lead-time AUROC in the same direction; the inversion result of §5.3 is therefore not specific to the within-issue pooling that defines C_{pool}.

3.5 Models

We evaluate three static survival architectures that span the classical-ML to deep-ML spectrum (random survival forests RSF [16], DeepSurv [18], and DeepHit [21]) and three sequence architectures trained end-to-end on per-issue trajectories (a vanilla LSTM, a Transformer encoder, and an S3-Hybrid LSTM that additionally consumes the five-dimensional H_{snap} summary at each time step via late fusion). The base Cox proportional-hazards model [8] and Kaplan–Meier estimator [17] provide diagnostic baselines. Implementations follow established Python packages: RSF via `scikit-survival` [27], DeepSurv/DeepHit via `pycox` [20], Cox PH/Kaplan–Meier via `lifelines` [9]. Hyperparameter search ranges, selected values, and full reproducibility scripts are provided in the companion repository https://github.com/kakeami/manga-survival-milets26.

4 Theory: Sufficient statistics for the two metrics

We formalise the central conceptual claim of the paper: pool concordance and lead-time AUROC at horizon k have different sufficient statistics under a hidden-Markov regime-switching DGP. Statements appear here; full self-contained proofs are in the companion long version (`long_proofs.pdf`, available in the companion repository alongside the reproducibility scripts of §3.5). We use $H_\tau = (X_1, \dots, X_\tau)$ for the history up to the evaluation landmark, $S_t(H) = P(T > t \mid H_t = H)$ for conditional survival, and $F_{\tau,k}(H_\tau) = P(T \le \tau + k \mid T > \tau, H_\tau)$ for the k-step cumulative event probability.

Theorem 1 (Pool-C_{pool} optimal score under stochastic ordering)**.** *Assume per-subject lifetimes are mutually independent given their histories. Suppose $\{S_t(H)\}_{t \ge \tau}$ is stochastically ordered by some scalar functional $\phi(H_\tau)$ (i.e., $\phi(H^{(1)}) \ge \phi(H^{(2)}) \Rightarrow S_t(H^{(1)}) \le S_t(H^{(2)})$ for all $t \ge \tau$). Then any strictly monotone transformation of ϕ achieves the supremum of pool concordance over measurable per-subject scores [30]. When stochastic ordering fails, the natural scalars $s_A = \langle \mathbf{h}_\tau^{\text{snap}}, \tilde{h} \rangle$ and $s_C(\cdot; k) = F_{\tau,k}$ generically rank histories differently on the regime simplex (Corollary 1); the pool-C_{pool}-attaining scalar then coincides with the rank function of the cohort-realised pair-sign $\operatorname{sgn}(\sigma_{ij} - 1/2)$, where $\sigma_{ij} = P(T_i < T_j \mid H_i, H_j) + \frac{1}{2} P(T_i = T_j \mid H_i, H_j)$ is the tie-corrected concordance, rather than with any of these natural scores, provided $\operatorname{sgn}(\sigma_{ij} - 1/2)$ is transitive on the cohort. When it is not (a Condorcet cycle), no per-subject scalar attains the unrestricted pairwise optimum $C^\star$ (the value achievable when each pair is predicted independently), so the per-subject supremum falls strictly below $C^\star$. Transitivity held in every sampled stationary HMM but is not established in general (companion* `long_proofs.pdf`, *§A.1).*

Lemma 1 (Lead-time-AUROC$_k$ optimal score)**.** *For any fixed horizon $k \ge 1$ and any landmark τ at which the prior $P(Y_k = 1 \mid T > \tau)$ is cohort-constant, the lead-time AUROC$_k$ is maximised over measurable per-subject scores by $s_k^*(H_\tau) = F_{\tau,k}(H_\tau)$, or any strictly monotone transformation thereof [25, 26].*

We work under the standard HMM assumptions [23, 28]: **(H1)** time-homogeneous regime chain with transition matrix $A_{l\ell} = P(s_{t+1} = \ell \mid s_t = l)$ and regime-conditional hazard $\tilde{h}(s) = P(T = t{+}1 \mid T > t, s_{t+1} = s)$ both t-independent; **(H2)** regime evolves before survival is resolved at each step, giving killed transition $\tilde{A}_{l\ell} = (1 - \tilde{h}(\ell))\, A_{l\ell}$; **(H3)** the observation X_t is conditionally independent of all other variables given s_t, so that, conditional on $\{T > \tau\}$, $P(T > \tau + k \mid T > \tau, s_\tau, H_\tau) = P(T > \tau + k \mid T > \tau, s_\tau)$ for all $k \ge 0$.

Theorem 2 (HMM bilinear factorisation)**.** *Under* (H1)–(H3)*, with regime $s_t \in \{1, \dots, K\}$, transition matrix A, and regime-dependent hazard $\tilde{h}(s)$, the k-step cumulative event probability factorises as*

$$F_{\tau,k}(H_\tau) \;=\; 1 - \left\langle \mathbf{h}_\tau^{\text{snap}},\; \mathbf{h}_\tau^{\text{traj}}(k) \right\rangle, \tag{1}$$

where $\mathbf{h}_\tau^{\text{snap}} = P(s_\tau \mid H_\tau, T > \tau)$ is the regime posterior at τ computed by the (survival-modified) HMM forward algorithm [28] and $\mathbf{h}_\tau^{\text{traj}}(k) = \tilde{A}^k \mathbf{1}$ is the k-iterate of the killed transition matrix $\tilde{A}_{l\ell} = (1 - \tilde{h}(\ell)) A_{l\ell}$ applied to $\mathbf{1}$.

Corollary 1 (Inversion under regime switching)**.** *Under the DGP of Theorem 2 with $K \ge 3$ regimes, the scores $s_A = \langle \pi, \tilde{h} \rangle$ and $s_C(\pi; k) = F_{\tau,k}(\pi)$ induce the same ranking on the regime simplex iff $\mathbf{h}_\tau^{\text{traj}}(k)$ lies in the affine span of $\tilde{h}$ and $\mathbf{1}$ with sign-coherent coefficient, i.e. $\tilde{A}^k \mathbf{1} = c\,\tilde{h} + d\,\mathbf{1}$ for some $c < 0, d \in \mathbb{R}$; for $k \ge 2$ this holds only on a Lebesgue-null subset of $(A, \tilde{h})$ (proved jointly in* `long_proofs.pdf` *§A.2), so the rankings are* distinct *for generic parameters. At $k = 1$ in the regime-persistent limit $A = (1{-}\epsilon)I + \epsilon M$, s_A and $s_C(\cdot; 1)$ coincide on the simplex up to $O(\epsilon)$, so s_A is first-order optimal for* AUROC$_1$*; alignment fails already at $\epsilon = 0$ once $k > 1$. By Lemma 1, no single per-subject scalar is simultaneously optimal for lead-time AUROC across all horizons, nor jointly optimal for both metrics. The claim is generic* non-equivalence *over the simplex; the* magnitude *of disagreement on a cohort (the negative rank correlation of §6) is empirical, witnessed by the K=3 construction of* `long_proofs.pdf` *§A.1.*

Theorems 1 and 2 and Lemma 1 thus tie the two factors of (1) to the feature families of §3.2: H_{snap} summarises $\mathbf{h}_\tau^{\text{snap}}$ and H_{traj} summarises $\mathbf{h}_\tau^{\text{traj}}(k)$, so their distinct empirical roles in §5 are predicted, not post hoc.

5 Main Results

We organise the empirical core around a single *apples-to-apples* grid: 6 architectures (LSTM, Transformer, S3-Hybrid, RSF, DeepSurv, DeepHit) × 7 feature families (MIN, BASE, R, H_{snap}, H_{traj}, H_{full}, HRP) × 5 prospective landmarks $\tau \in \{1990, \ldots, 2010\}$, with undefined combinations omitted (static models cannot accept the variable-length MIN; S3-Hybrid is restricted to {MIN, H_{snap}, H_{full}, HRP}). This yields 36 defined cells per landmark (= 7+7+4+6+6+6): a 180-cell pool-concordance grid across the five landmarks and an 18-cell sequence subgrid per landmark (LSTM, Transformer, S3-Hybrid) for lead-time AUROC at horizons $k \in \{1, \ldots, 26\}$. Three observations follow.

5.1 Pool concordance: raw current state is near-ceiling

Across the 5-landmark, 5-seed mean of C_{pool}, the matrix top is **LSTM × MIN** at $C_{\text{pool}} = 0.791$: a vanilla LSTM on the raw three-column trajectory (`rank_in_issue`, `pages`, `magazine`) with no engineered features (Table 1a). The strongest static configuration, DeepSurv × H_{full}, reaches $C_{\text{pool}} = 0.757$, and the strongest engineered-feature sequence cell, LSTM × H_{traj}, reaches 0.770. The engineering gradient under LSTM (MIN → H_{snap} → HRP, i.e. from raw input to widest engineered family) is in fact *reversed*: adding engineered features *degrades* pool concordance from 0.791 to 0.752, with the gap widening at the latest landmark (LSTM × H_{full} at τ=2010 is 0.133 below LSTM × MIN). The HMM family itself is informative when isolated: under every static architecture (DeepSurv, RSF, DeepHit), the snapshot–trajectory contrast $\Delta_{H_{\text{snap}} - H_{\text{traj}}}$ exceeds +0.075, showing that the latest latent state carries dominant signal relative to its trajectory aggregates.

5.2 Lead-time AUROC: the picture inverts under multi-step forecasting

Under lead-time AUROC the ranking inverts. The clearest signature is visible in Table 1: the per-row maxima of the pool C_{pool} matrix (1a) and the lead-time $\text{AUROC}_k@k$=8 matrix (1b) sit in different feature columns for every architecture, and the architectures that top the pool matrix (LSTM with MIN or H_{snap}) are displaced by sequence-with-engineered-features cells (LSTM/Transformer with H_{full} or HRP) in the lead-time matrix. **LSTM** × H_{full} leads the sequence-cell matrix at every horizon $k \in \{1, \ldots, 26\}$, from $\text{AUROC}_k = 0.849$ at k=1 to 0.807 at k=26 (Table 2); at k=1 alone, Transformer × H_{full} near-ties at 0.850 (a 0.001 gap, well within seed noise; §7.1), and LSTM × H_{full} is strictly top at every $k \geq 2$. The gap to the pool-concordance champion LSTM × MIN is +0.065 at k=1 and widens to +0.074 at k=26; under a same-snapshot paired bootstrap with B=500 resamples, the per-cell CI is bounded away from zero at every k. A direct head-to-head against the static DeepSurv × H_{snap} configuration is even more decisive: the paired difference $\Delta_k^\tau = \text{AUROC}_k^{\text{LSTM}\times H_{\text{full}}} - \text{AUROC}_k^{\text{DS}\times H_{\text{snap}}}$ is positive in 130/130 cells (5 landmarks × 26 horizons), with range [+0.106, +0.179] and CI lower bound strictly above zero everywhere (Figure 2). All paired CIs in this subsection are computed across 5-seed means; per-seed dispersion and the strict-threshold replication verdict are reported in §7.1. Static cells never enter the lead-time top-5: their score at landmark τ is constant in $t > \tau$, whereas sequence models update at every issue and produce horizon-dependent rankings.

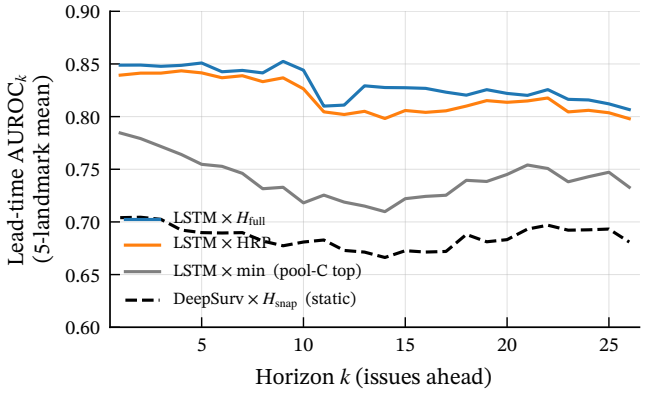

Figure 2: Lead-time AUROC at horizon k, mean across the five landmarks. Pool-concordance champion LSTM × MIN (grey) is exceeded at every k by LSTM × H_{full} (blue) and LSTM × HRP (orange). The static DeepSurv × H_{snap} reference (dashed black) is the lowest of the four; static models are dominated at every horizon under matched bootstrap.

The two metrics thus reward orthogonal feature representations on the same data, architectures, and training pool: raw trajectory features suffice for pool concordance, while lead-time AUROC at horizon k requires *detail-preserving* features summarising regime dynamics.

5.3 Inversion universality across feature classes

The dissociation above is not specific to the HMM family. We test *feature-class invariance* by training LSTM with four alternative feature classes on the same protocol: lag-shifted `placement_ma4` ($k \in \{3, 6, 12\}$, three cells), exponentially weighted moving averages of `placement_ma4` ($\alpha \in \{0.1, 0.3, 0.5\}$, three cells), and Bayesian online change-point statistics following Adams and MacKay [1] (one cell). Combined with the seven LSTM cells of §5.1 (MIN, BASE, R, H_{snap}, H_{traj}, H_{full}, HRP), this gives 14 LSTM cells per landmark.

For each landmark τ, we rank the 14 cells by pool C_{pool} and separately by lead-time AUROC_k at k=8, then compute the Spearman rank correlation between the two orderings (Table 3). The correlation is uniformly negative across all five landmarks: $\rho \in [-0.79, -0.52]$, mean −0.67. The top-3 cells of the two rankings are *disjoint* (zero overlap) at every landmark. Pool-concordance top cells are simple or aggregating (MIN, H_{snap}, CP, EWMA-0.5), while lead-time top cells are detail-preserving (LAG-12, H_{full}, HRP, LAG-3).

We interpret this as evidence that the pool-vs-lead-time inversion is a property of how the feature class summarises trajectory dynamics, not of any single feature family. HMM, lag, EWMA, and

Table 1: Apples-to-apples matrix view of the metric pair, with the same architecture×feature layout on both sides. The two matrices share row/column ordering so the reader can compare any cell directly: bolded entries are per-row maxima, and the bolded cells sit in different columns in (a) and (b) for every architecture (e.g. for DeepSurv, H_{full} wins pool C_{pool} while H_{snap} wins LT-AUROC). Dashes mark architecture/feature pairs that are undefined (static models do not accept the raw MIN trajectory; the S3-Hybrid architecture is evaluated on a subset of families).

(a) Pool concordance C_{pool} (5-landmark mean).

arch	MIN	BASE	R	H_{snap}	H_{traj}	H_{full}	HRP
LSTM	**0.791**	0.758	0.767	0.769	0.770	0.729	0.752
Transformer	0.736	0.757	0.727	**0.761**	0.740	0.753	0.739
S3-Hybrid	**0.785**	—	—	0.692	—	0.759	0.746
DeepSurv	—	0.653	0.688	0.741	0.661	**0.757**	0.749
RSF	—	0.588	0.680	0.741	0.644	0.734	**0.747**
DeepHit	—	0.559	0.659	0.723	0.600	**0.754**	0.752

(b) Lead-time AUROC at k=8 (5-landmark mean).

arch	MIN	BASE	R	H_{snap}	H_{traj}	H_{full}	HRP
LSTM	0.732	0.724	0.718	0.721	0.770	**0.841**	0.833
Transformer	0.787	0.807	0.827	0.809	0.813	**0.835**	0.829
S3-Hybrid	0.731	—	—	**0.836**	—	0.769	0.817
DeepSurv	—	0.495	0.620	**0.682**	0.461	0.670	0.651
RSF	—	0.627	0.647	**0.721**	0.669	0.699	0.712
DeepHit	—	0.573	0.577	0.670	0.526	0.681	**0.697**

Table 2: Lead-time AUROC at representative horizons for the six cells with highest mean AUROC_k across k. Bold marks the top cell at each horizon. LSTM × H_{full} is the matrix top at every $k \geq 2$ and near-ties Transformer × H_{full} at k=1 ($\Delta = 0.001$, within seed noise).

arch	feat	k=1	k=4	k=8	k=16	k=26
LSTM	H_{full}	0.849	**0.849**	**0.841**	**0.827**	**0.807**
Transformer	H_{full}	**0.850**	0.845	0.835	0.824	0.801
Transformer	HRP	0.835	0.831	0.829	0.821	0.802
S3-Hybrid	H_{snap}	0.836	0.837	0.836	0.815	0.787
Transformer	R	0.835	0.832	0.827	0.814	0.791
LSTM	HRP	0.839	0.843	0.833	0.804	0.798
LSTM ref.	*MIN*	*0.784*	*0.764*	*0.732*	*0.724*	*0.733*

Table 3: Inversion universality across feature classes. Each row is one landmark; TOP-3 POOL and TOP-3 LEAD-TIME (k=8) list the three highest-ranked of 14 LSTM cells under each metric. OVERLAP counts shared cells (always 0). Spearman ρ is computed over all 14 cells.

τ	ρ	TOP-3 POOL	TOP-3 LEAD-TIME (k=8)	OVERLAP
1990	−0.79	{MIN, H_{snap}, BASE}	{LAG-12, H_{full}, R}	0/3
1995	−0.64	{MIN, CP, H_{snap} }	{LAG-12, HRP, H_{full}}	0/3
2000	−0.64	{EWMA-0.5, CP, BASE}	{LAG-3, LAG-12, H_{full}}	0/3
2005	−0.52	{CP, MIN, EWMA-0.5}	{LAG-3, LAG-12, H_{full}}	0/3
2010	−0.76	{R, CP, H_{snap} }	{HRP, H_{full}, LAG-3}	0/3
mean	**−0.67**	—	—	**0/3**

BOCPD are four mechanistically different ways to encode placement and regime information, and all four reproduce the same task-dependent ranking under our protocol.

§5.2–§5.3 refute the natural reading of §5.1 that "raw features suffice": the pool-winning cell (LSTM × MIN) is *never* the lead-time-winning cell, and lead-time top cells uniformly use engineered features (HMM, lag, EWMA, or HRP). Theoretical interpretation appears in §6, seed-level robustness (per-seed dispersion and the full replication verdict for the 5-seed means reported above) in §7.

6 Theory-empirical link

§5 established the inversion empirically across four feature classes. We connect it to the bilinear factorisation of Theorem 2–Corollary 1.

6.1 Cell-level inversion at the metric-pair level

Theorem 1 and Lemma 1 predict that pool concordance and lead-time AUROC_k admit different optima whenever conditional survival curves cross; Corollary 1 sharpens this to a guarantee of inversion under HMM regime switching with $K \geq 3$ (the K=2 case, where inversion can occur for anti-persistent transitions but not horizon-dependently, is degenerate and we report it empirically below as a robustness check). The natural cell-level signature is a negative rank correlation between the two metrics across the architecture–feature grid. Across the 18 sequence cells of §5, the Spearman rank correlation between cells' pool C_{pool} rank and their lead-time AUROC_k rank is negative at every horizon (Table 4): ρ ranges from −0.637 at k=1 to a minimum of −0.769 at k=8 and k=13, with two-sided $p < 0.005$ at every k. The top-ranked cell under the two metrics disagrees at all seven horizons we report. The inversion is therefore not concentrated at a single k: it is a property of the metric pair, consistent with Corollary 1.

The same negative rank correlation replicates along three independent axes that probe distinct sources of variability. *(i) Seed.* Across all 25 (seed, τ) pairs of the 550-fit replication grid (§7.1), Spearman ρ between pool C_{pool} and lead-time AUROC_k at k=8 over the 14 LSTM universality cells is negative in all 25/25 configurations; the corresponding binomial null probability of obtaining 25/25 negatives by chance is 2.98×10^{-8}. *(ii) HMM state count.* Refitting the 11 HMM-dependent cells at each $K \in \{2, 3, 5, 6, 7, 8\}$ and recomputing the rank correlation gives negative ρ in all 35/35 (K, τ) configurations (§7.2). *(iii) Concordance metric.* Replacing pool C_{pool} with Antolini's IPCW-weighted work-level C-index gives negative ρ in 5/5 landmarks and in 99/100 (τ, horizon) combinations. The empirical inversion is therefore not a seed, state-count, or pooling artefact: it is a metric-pair property of the kind Corollary 1 predicts under crossing-hazard regime switching.

6.2 Information-theoretic corroboration

Equation (1) predicts that the trajectory factor $\mathbf{h}_\tau^{\text{traj}}(k)$ is essential for long-horizon prediction, while the snapshot factor $\mathbf{h}_\tau^{\text{snap}}$

Table 4: Cell-level rank correlation between pool concordance and lead-time AUROC$_k$ across the 18 sequence cells of §5, at seven horizons. Negative ρ confirms Corollary 1; top cells under the two metrics differ at every k.

k	ρ	p	top-1 pool C_{pool}	top-1 AUROC$_k$
1	−0.637	0.004	LSTM × MIN	Transformer × H_{full}
2	−0.666	0.003	LSTM × MIN	Transformer × H_{full}
4	−0.736	0.0005	LSTM × MIN	LSTM × H_{full}
8	−0.769	0.0002	LSTM × MIN	LSTM × H_{full}
13	−0.769	0.0002	LSTM × MIN	Transformer × H_{full}
18	−0.719	0.0008	LSTM × MIN	LSTM × H_{full}
26	−0.672	0.002	LSTM × MIN	LSTM × H_{full}

alone suffices when k is short. We test this directly using mutual information. Let $Y_k = \mathbb{1}\{T \le \tau + k\}$ be the lead-time label and $\hat{I}$ a KSG-style mutual information estimator [29]. The conditional contributions $\hat{I}(H_{\text{traj}}; Y_k \mid H_{\text{snap}})$ and $\hat{I}(H_{\text{snap}}; Y_k \mid H_{\text{traj}})$, averaged over the five landmarks, move in the directions Theorem 2 predicts: $\hat{I}(H_{\text{traj}}; Y_k \mid H_{\text{snap}})$ grows monotonically with k and is bounded away from zero with 95 % bootstrap CI at every $k \ge 4$ in 5/5 landmarks, while $\hat{I}(H_{\text{snap}}; Y_k \mid H_{\text{traj}})$ is approximately zero at k=1 and becomes non-positive at $k \ge 13$ as a point estimate.[1] The ratio of marginal estimates $\hat{I}(H_{\text{traj}}; Y_k)/\hat{I}(H_{\text{snap}}; Y_k)$ rises from 0.91 at k=1 to 1.47 at k=26: at short k the snapshot factor dominates the trajectory; at long k the trajectory dominates the snapshot, the empirical signature of (1). The corresponding network-level ablation, a counterfactual deletion of the H_{traj} features inside a trained LSTM, is reported in §7.3; the direction of effect matches the population prediction here, while the paired confidence intervals do not separate strictly from zero. We read the population-level information-theoretic result in this subsection as the principal test of the theorem, and the deep-model ablation as a single-network sanity check.

6.3 Synthetic regime-switching simulation

To verify that the inversion is intrinsic to crossing-hazards regime dynamics rather than to a peculiarity of the manga cohort, we simulate two contrasting DGPs with K=3 regimes. A *monotone* DGP with hazards $\tilde{h} = (0.05, 0.15, 0.40)$ and approximately uniform mixing satisfies the stochastic-ordering hypothesis of Theorem 1; here the oracle current-hazard and oracle $F_{\tau,k}$ score have Spearman $\rho \approx 1.00$ at every k and no inversion appears. A *crossing-hazards* DGP with $\tilde{h} = (0.10, 0.05, 0.001)$ and asymmetric transitions violates stochastic ordering; the same two oracle scores decouple as ρ falls from 1.000 at k=1 to 0.537 at k=10 over 5 seeds, and a current-state proxy LSTM's correlation with the oracle $F_{\tau,k}$ collapses from 0.91 to 0.47 over the same range. The cell-level inversion observed on real data ($\rho \approx -0.71$) is therefore a strong-form realisation of the same phenomenon that the theorem identifies in the smaller synthetic system.

[1]The KSG chain rule can yield mildly negative estimates when X is informationally subsumed by Z [29]; the sign here is consistent both with that small-sample behaviour and with Theorem 2 once k exceeds the regime mixing time.

7 Replication and limitations

7.1 Five-seed replication of the inversion

We replicated the 110-cell LSTM fit grid that underpins the inversion claim (75 sequence cells of §5 plus 35 lag, EWMA, and BOCPD baseline cells) across five independent training seeds, for a total of 550 fits, and recomputed all per-cell pool C_{pool} and lead-time AUROC$_k$ values. The replication targets four pre-specified checks (Table 5); we read **R2** as the central universality check and **R1** as the companion direction check. **R2**, the feature-class inversion universality (§5.3), formalised as "Spearman $\rho(C_{\text{pool}}, \text{AUROC}_k@k{=}8)$ over 14 LSTM cells is negative," must hold across all (seed, τ) configurations. It does, in all 25/25 configurations: $\rho \in [-0.877, -0.055]$ with median $\rho \approx -0.65$. **R1** asks whether LSTM × H_{full} exceeds LSTM × MIN on lead-time AUROC at sufficiently many cells; the mean direction holds across all five seeds (mean $\Delta \in [+0.015, +0.045]$, 5/5 seeds positive), while the pre-registered strict threshold ($\ge 14/26$ horizons at $\ge 4/5$ landmarks) passes in only 2/5 seeds. We interpret this as evidence that the 14-horizon threshold sits within seed-level noise even though the direction is uniform; the universality verdict (R2) does not depend on R1's strict threshold.

Per-cell AUROC dispersion across seeds has median 0.027 (single seed) and ≈ 0.013 (5-seed mean). Pool C_{pool} cell rankings are seed-sensitive among cells of similar performance (median across-pair Kendall $\tau \in [0.18, 0.27]$) but this does not perturb the inversion direction at any landmark.

7.2 State-count robustness

The choice K=4 in §3.2 could in principle drive the empirical inversion through a particular quantisation of placement regimes. To rule this out, we refit the HMM at each $K \in \{2, 3, 5, 6, 7, 8\}$ on the same training pool and recomputed the lead-time AUROC grid for the 11 HMM-dependent cells (LSTM, Transformer, and S3-Hybrid each paired with $H_{\text{snap}}, H_{\text{traj}}, H_{\text{full}}$, HRP), keeping all other hyperparameters fixed. The pool-vs-lead-time inversion, defined as the five-landmark mean Spearman $\rho(C_{\text{pool}}, \text{AUROC}_k@k{=}8)$ over the 11 cells, is negative at every K, with ρ in $[-0.86, -0.67]$ and all 35 (K, τ) configurations negative. Per-cell AUROC$_k@k$=8 varies by at most 0.044 between K=2 and K=8 (mean 0.016), so the cells' *absolute* discrimination is nearly K-invariant; only the cells' *ranking* varies with K, and that ranking inverts in the same direction at every K. We therefore retain K=4 in the main results without further qualification on state-count sensitivity.

7.3 Counterfactual deletion and remaining limitations

A complementary sanity check ablates each HMM factor in a single LSTM × H_{full} trained model by mean-imputing five H_{snap} or five H_{traj} columns at inference time, leaving weights and architecture untouched, and measuring the lead-time AUROC drop Δ_{snap} and Δ_{traj}. The direction of effect matches the trajectory-factor narrative in 4/5 landmarks: $\Delta_{\text{snap}} > 0$, $\Delta_{\text{traj}} > 0$, and the ratio $\Delta_{\text{traj}}/|\Delta_{\text{snap}}|$ grows with k. The paired CI bound, however, separates from zero in 0/5 landmarks for $k \ge 8$, and at τ=1995 the trajectory ablation *improves* lead-time AUROC by approximately 0.003 over the unmasked baseline. The deep model therefore qualitatively but not

Table 5: Five-seed replication: pre-registered checks (R1–R4) on the 550-fit grid. R2 is the central robustness check (REQUIRED PASS); R1 is required but interpreted via mean direction rather than strict horizon threshold; R3 and R4 are nice-to-have characterisations.

check	target	result	verdict
R1	LSTM×H_{full} >MIN on AUROC$_k$, $\geq 14/26$ k at $\geq 4/5$ τ	2/5 seeds strict; mean $\Delta > 0$ in 5/5 seeds	direction PASS
R2	$\rho(C_{\text{pool}}, \text{AUROC}_k@k{=}8) < 0$ in all (seed, τ) over 14 cells	25/25 negative, median $\rho \approx -0.65$	**PASS**
R3	per-(cell, τ, k) seed std < 0.02 in $\geq 90\%$ cells	40.8%; median std 0.027	report only
R4	per-τ pool C_{pool} ranking Kendall $\tau \geq 0.7$	median $[0.18, 0.27]$	report only

strictly reproduces the unconditional information-theoretic direction of §6.2; we interpret the τ=1995 idiosyncrasy as evidence that Corollary 1 is a population statement about sufficient statistics, not a guarantee about the internal weighting of any one trained network. The unconditional inversion claim itself is carried by the cell-level rank correlation (§6.1) and the KSG conditional MI estimates (§6.2); the ablation here is a single-network diagnostic and not a test of the theorem.

We highlight six remaining limitations that scope the strength of the conclusions. **(L1) Theorem 1 requires stochastic ordering.** This is an assumption on the conditional survival family, not an artefact of regime-switching; it is approximately satisfied in our corpus by regime persistence but need not hold on adjacent serial publication formats. The synthetic crossing-hazards DGP of §6.3 models the violation regime explicitly: when stochastic ordering fails, the oracle current-hazard and oracle $F_{\tau,k}$ scores decouple and the empirical inversion emerges in the same direction as on the manga cohort. Formats-portability is therefore an empirical question about other cohorts, not a gap in the theory itself. **(L2) Calibration is evaluated via a score-rescaling proxy.** We rescale each model's per-cell output to $[0, 1]$ at each landmark and horizon and treat it as a probability for integrated-Brier scoring, because reconstructing the discrete-time hazard $S(t \mid x)$ from saved checkpoints lies outside the workshop scope. The proxy supports only *relative* comparison across cells, not absolute calibration. Under this proxy the top five lead-time cells achieve a five-landmark mean integrated Brier of 0.31 against 0.45 for the bottom five cells, indicating that the lead-time-best configurations are not systematically miscalibrated relative to their peers. Because the proxy certifies only relative ordering, any *absolute* deployment decision (e.g., thresholding a reader-facing cancellation probability) is unsupported until the discrete-time hazard $S(t \mid x)$ is reconstructed; that full integrated-Brier comparison is deferred to a long companion version. **(L3) Causal effect of editorial response is out of scope, and reader-side signal is unobserved.** The protocol evaluates predictive discrimination, not the intervention effect that a deployed forecast might trigger via reader or editorial behaviour; the MADB corpus surfaces only the placement realisation, so a shift in reader preference and an editorial intervention applied to a fixed reader signal are jointly absorbed into the latent regime state. A deployment study with controlled-rollout infrastructure is a natural follow-on. **(L4) External validity beyond Japanese weekly manga.** Our empirical claims are conditioned on a single cultural–editorial system: four weekly *shōnen* magazines with a reader-questionnaire feedback loop that is largely Japanese in form. The theory (Theorem 2, Corollary 1) is domain-agnostic and a stylised crossing-hazards DGP reproduces the inversion (§6.3), but the magnitude and timing of the inversion on adjacent serialised formats (Western comics, web-novel platforms, streaming-series renewals) remain an empirical question requiring cohort-specific replication. We therefore frame the contribution as single-domain evidence for a domain-agnostic mechanism rather than as a cross-domain empirical claim. **(L5) The cancellation label conflates ending types.** Our operational definition (the last issue in which a work appears, §3.1) pools editorially forced cancellation with planned endings, indefinite hiatuses, and occasional MADB record artefacts; the corpus carries no editorial reason code, so this label noise is not separable here. It is plausibly non-differential across feature cells (every model sees the same labels) and so is unlikely to drive the inversion, but it caps the attainable absolute discrimination. **(L6) Lead-time AUROC$_k$ is not censoring-corrected.** The label $Y_k = \mathbf{1}[\tau < T \leq \tau + k]$ is scored without inverse-probability-of-censoring weighting. Administrative censoring is confined to the final window year (§3.1), so few test-cohort works are censored within $k \leq 26$ and the metric is at most mildly optimistic; the pool side already reports an IPCW-weighted Antolini check (§6.1), so an IPCW-weighted lead-time AUROC is a direct extension.

8 Discussion

The central observation of this paper is that two natural deployment metrics for survival forecasting on long-horizon editorial time series do not rank features in the same order. On the manga cohort, the strongest cells on lead-time AUROC at every horizon $k \in \{1, \ldots, 26\}$ are sequence models with engineered regime-encoding features (H_{snap}/H_{traj}, lag-k, EWMA-α, or BOCPD); the strongest cell on issue-level pool concordance is a LSTM on raw three-column placement trajectories alone. The dissociation is feature-class invariant and seed-robust, and is predicted by an HMM regime-switching argument in which lead-time AUROC$_k$ admits a bilinear sufficient-statistic decomposition $F_{\tau,k} = 1 - \langle \mathbf{h}^{\text{snap}}, \mathbf{h}^{\text{traj}}(k) \rangle$ while pool concordance is solved by the snapshot factor alone under approximate stochastic ordering.

For the reader-facing deployment scenario motivated in §1, the appropriate feature representation is the lead-time-AUROC-optimal cell: LSTM/Transformer with the trajectory factor encoded explicitly (efficiently but not uniquely by HMM, since lag-k, EWMA, and BOCPD families reproduce the same horizon-dependent advantage). The complementary editorial use case (single-issue ranking) is solved near its ceiling by raw placement trajectories on a vanilla LSTM. The metric and the deployment use case are therefore not

separable in practice: a configuration tuned for one will, by Corollary 1, generically fail the other.

Future work. A long companion version pursues three extensions left to scoping here: a discrete-time-hazard calibration study (reconstructing $S(t \mid x)$ rather than the score-rescaling proxy of L2), an issue-level Antolini analysis, and per-architecture SHAP attributions linking each factor of (1) to specific features. Our architecture grid is deliberately survival-native (LSTM, Transformer, S3-Hybrid, RSF, DeepSurv, DeepHit); benchmarking recent general-purpose deep forecasting backbones (e.g., MLF [33], TimePro [22], and SEMixer [34]) as trajectory encoders, alongside mainstream feature-selection baselines, is a complementary direction we leave open. The deployment implications remain an open question that an archival protocol cannot settle and a controlled-rollout study would test. In particular, whether a lead-time-AUROC-optimal forecast surfaced to readers actually shifts questionnaire-response behaviour enough to alter the cancellation hazard it predicts is a causal question, not a discrimination one.

Acknowledgments

Claude (Anthropic) was used to assist with language editing and manuscript preparation. The author reviewed and verified all content and is solely responsible for the paper.

References

[1] Ryan Prescott Adams and David J. C. MacKay. 2007. Bayesian Online Changepoint Detection. arXiv:0710.3742 [stat.ML] https://arxiv.org/abs/0710.3742

[2] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. 2005. A Time-Dependent Discrimination Index for Survival Data. *Statistics in Medicine* 24, 24 (2005), 3927–3944. doi:10.1002/sim.2427

[3] Christoph Bergmeir and José M. Benítez. 2012. On the Use of Cross-Validation for Time Series Predictor Evaluation. *Information Sciences* 191 (2012), 192–213. doi:10.1016/j.ins.2011.12.028

[4] Christoph Bergmeir, Rob J. Hyndman, and Bonsoo Koo. 2018. A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction. *Computational Statistics & Data Analysis* 120 (2018), 70–83. doi:10.1016/j.csda.2017.11.003

[5] Sudip Bhattacharjee, Ram D. Gopal, Kaveepan Lertwachara, James R. Marsden, and Rahul Telang. 2007. The Effect of Digital Sharing Technologies on Music Markets: A Survival Analysis of Albums on Ranking Charts. *Management Science* 53, 9 (2007), 1359–1374. doi:10.1287/mnsc.1070.0699

[6] Eric T. Bradlow and Peter S. Fader. 2001. A Bayesian Lifetime Model for the "Hot 100" Billboard Songs. *J. Amer. Statist. Assoc.* 96, 454 (2001), 368–381. doi:10.1198/016214501753168091

[7] Cristian Candia, C. Jara-Figueroa, Carlos Rodriguez-Sickert, Albert-László Barabási, and César A. Hidalgo. 2019. The Universal Decay of Collective Memory and Attention. *Nature Human Behaviour* 3 (2019), 82–91. doi:10.1038/s41562-018-0474-5

[8] David R. Cox. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34, 2 (1972), 187–202. doi:10.1111/j.2517-6161.1972.tb00899.x

[9] Cameron Davidson-Pilon. 2019. lifelines: Survival Analysis in Python. *Journal of Open Source Software* 4, 40 (2019), 1317. doi:10.21105/joss.01317

[10] Arthur De Vany and W. David Walls. 1999. Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office? *Journal of Cultural Economics* 23, 4 (1999), 285–318. doi:10.1023/A:1007608125988

[11] Sean R. Eddy. 1996. Hidden Markov Models. *Current Opinion in Structural Biology* 6, 3 (1996), 361–365. doi:10.1016/S0959-440X(96)80056-X

[12] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. 2011. A Sticky HDP-HMM with Application to Speaker Diarization. *Annals of Applied Statistics* 5, 2A (2011), 1020–1056. doi:10.1214/10-AOAS395

[13] Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. 1996. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine* 15, 4 (1996), 361–387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

[14] Patrick J. Heagerty and Yingye Zheng. 2005. Survival Model Predictive Accuracy and ROC Curves. *Biometrics* 61, 1 (2005), 92–105. doi:10.1111/j.0006-341X.2005.030814.x

[15] Rob J. Hyndman and George Athanasopoulos. 2018. *Forecasting: Principles and Practice* (2nd ed.). OTexts, Melbourne, Australia. https://otexts.com/fpp2/

[16] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. Random Survival Forests. *Annals of Applied Statistics* 2, 3 (2008), 841–860. doi:10.1214/08-AOAS169

[17] E. L. Kaplan and Paul Meier. 1958. Nonparametric Estimation from Incomplete Observations. *J. Amer. Statist. Assoc.* 53, 282 (1958), 457–481. doi:10.1080/01621459.1958.10501452

[18] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. 2018. DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology* 18 (2018), 24. doi:10.1186/s12874-018-0482-1

[19] Sharon Kinsella. 2000. *Adult Manga: Culture and Power in Contemporary Japanese Society.* Curzon Press, Richmond, Surrey.

[20] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. 2019. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research* 20, 129 (2019), 1–30. https://jmlr.org/papers/v20/18-424.html

[21] Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. 2018. DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence.* AAAI Press, New Orleans, Louisiana, USA, 2314–2321. doi:10.1609/aaai.v32i1.11842

[22] Xiaowen Ma, Zhen-Liang Ni, Shuai Xiao, and Xinghao Chen. 2025. TimePro: Efficient Multivariate Long-term Time Series Forecasting with Variable- and Time-Aware Hyper-state. In *Proceedings of the 42nd International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 267).* PMLR, 42096–42111. https://proceedings.mlr.press/v267/ma25p.html

[23] Kevin P. Murphy. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning.* Ph. D. Dissertation. University of California, Berkeley. https://www.cs.ubc.ca/~murphyk/Thesis/thesis.html

[24] National Center for Art Research. 2024. Media Arts Database. Independent Administrative Institution National Museum of Art, Japan. https://mediaarts-db.artmuseums.go.jp Officially launched 2024-01-31; made freely available for reuse under the database's own terms of use (source and version attribution requested). Accessed 2026-05-21..

[25] Jerzy Neyman and Egon S. Pearson. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A* 231 (1933), 289–337. doi:10.1098/rsta.1933.0009

[26] Margaret Sullivan Pepe. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press.

[27] Sebastian Pölsterl. 2020. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research* 21, 212 (2020), 1–6. https://www.jmlr.org/papers/v21/20-729.html

[28] Lawrence R. Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* 77, 2 (1989), 257–286. doi:10.1109/5.18626

[29] Brian C. Ross. 2014. Mutual Information between Discrete and Continuous Data Sets. *PLoS ONE* 9, 2 (2014), e87357. doi:10.1371/journal.pone.0087357

[30] Moshe Shaked and J. George Shanthikumar. 2007. *Stochastic Orders.* Springer. doi:10.1007/978-0-387-34675-5

[31] Marc Steinberg. 2012. *Anime's Media Mix: Franchising Toys and Characters in Japan.* University of Minnesota Press, Minneapolis.

[32] Leonard J. Tashman. 2000. Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review. *International Journal of Forecasting* 16, 4 (2000), 437–450. doi:10.1016/S0169-2070(00)00065-0

[33] Xu Zhang, Zhengang Huang, Yunzhi Wu, Xun Lu, Erpeng Qi, Yunkai Chen, Zhongya Xue, Qitong Wang, Peng Wang, and Wei Wang. 2025. Multi-period Learning for Financial Time Series Forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25).* ACM, 2848–2859. doi:10.1145/3690624.3709422

[34] Xu Zhang, Qitong Wang, Peng Wang, and Wei Wang. 2026. SEMixer: Semantics Enhanced MLP-Mixer for Multiscale Mixing and Long-term Time Series Forecasting. In *Proceedings of the ACM Web Conference 2026 (WWW '26).* ACM, 5636–5647. doi:10.1145/3774904.3792639