SPARSE MULTI-LANDMARK SUPERVISION FOR LONGITUDINAL EHR SURVIVAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electronic health records (EHRs) accumulate irregularly over years, yet survival models that support longitudinal risk prediction are commonly trained at only one prediction time per patient and then queried repeatedly as new clinical observations arrive. We study sparse multi-landmark supervision, which trains from several points along each patient trajectory, as a model- and objective-agnostic strategy to reduce this mismatch. Across eight tasks, two health systems, and six model families, training with three landmarks improves integrated Brier score in all 48 task–model configurations and concordance index in 45 of 48 under multi-landmark evaluation. These gains persist when optimizer updates or input tokens are matched, when recent clinical history is progressively truncated, and across retraining. Multi-landmark supervision also weakens coupling between predicted risk and EHR recording patterns (e.g., code density and redundancy). On EHRSHOT, three-landmark training can further compensate for substantially fewer training patients. Beyond predictive performance, we introduce trajectory-aware evaluation for models used repeatedly over time. Multi-landmark supervision makes successive predictions more temporally consistent, while direct trajectory analysis reveals that consistency and informative risk evolution are distinct properties. A ranking-augmented proof of principle shows that explicitly relating predictions across landmarks can alter trajectory direction. In a nutshell, our results show that better use of existing longitudinal records can substantially improve survival prediction, while motivating models and evaluation that explicitly account for how risk evolves within patients over time.

1 INTRODUCTION

Electronic health records (EHRs) provide a longitudinal but incomplete view of patient health, often spanning years of clinical care. They can combine diagnoses, procedures, laboratory measurements, medications, notes, and other modalities, creating opportunities to model disease progression and evolving health risks (Placido et al., 2023; Shmatko et al., 2025). Effective risk prediction, however, depends not only on how these heterogeneous records are represented, but also on how they are repeatedly evaluated for in-patient arrivals.

In deployment, a survival model may be queried whenever a patient returns, from years before an outcome to shortly before it. Training, however, typically uses a single prediction time per patient, often the latest eligible point before the event or end of follow-up. This creates a mismatch between how models are trained and how they are used. In survival analysis, the prediction time determines both the observed history and the residual time to the outcome; training only at the latest landmark therefore covers a narrow subset of the prediction contexts encountered during follow-up. This mismatch is particularly important in EHRs, where records are irregular and bursty: code density and repetition vary with clinical activity, and similar input lengths can correspond to very different spans of calendar time and clinical context (Wornow et al., 2025; Chen et al., 2026).

Recent EHR models have increasingly borrowed from natural-language modeling, using large sequence and foundation models trained on millions of patients and billions of clinical events (Zhang et al., 2026; Gadd et al., 2026). However, EHRs are long, heterogeneous token sequences with irregular temporal processes, where the spacing and timing of observations carry clinical information. EHR modeling therefore sits between language modeling and time-series analysis. Scaling

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.

model capacity does not by itself address the temporal and recording-specific structure of EHRs, or how repeated risk predictions should be evaluated for robustness. This motivates complementary work on how longitudinal supervision is constructed and how models are assessed beyond aggregate predictive performance.

Existing approaches address the temporal structure of EHRs mainly through representation or architecture. For instance, temporal encodings such as CT-ROPE (Li et al., 2025) and "no-event" padding (Shmatko et al., 2025) incorporate irregular inter-event gaps, while dynamic survival models evolve latent patient states or regularize predictions across time (Bleistein et al., 2024; Zeng et al., 2025; Fah et al., 2026). We instead ask a simpler question about the learning signal itself: at which points along a patient's trajectory should survival prediction be supervised? Classical landmarking re-baselines prediction at multiple times while the patient remains at risk (Houwelingen, 2007; Tanner et al., 2021). We adapt this idea as sparse multi-landmark supervision for deep EHR survival models.

We ask whether training at several prediction times per patient provides a model- and objective-agnostic improvement to survival prediction. For each patient, we sample a small set of eligible landmarks, construct the history available at each, and re-baseline the residual event or censoring time. This sparse multi-landmark supervision broadens the range of patient states seen during training without requiring additional patients. We evaluate it across eight tasks from EHRSHOT (Wornow et al., 2023a) and MIMIC-IV (Johnson et al., 2024), six model families, and matched-budget controls.

But broader supervision raises a second question: when the same patient is scored repeatedly, do better snapshot predictions also form better trajectories? Standard concordance and Brier scores evaluate predictions individually, not how they evolve across visits. We therefore extend evaluation to robustness under record perturbations and retraining, temporal consistency, hazard variation, and patient-level risk direction. This distinction is central to our findings: multi-landmark supervision improves prediction and makes successive estimates more consistent, yet consistency alone does not guarantee informative temporal dynamics.

Our contributions are four-fold:

- **Sparse multi-landmark supervision improves EHR survival prediction across models and objectives.** Across eight tasks, two health systems, and six model families, the gain persists under matched optimizer and token budgets. On EHRSHOT, sparse landmarks can also partly compensate for fewer training patients.
- **The gains extend beyond average predictive performance.** Multi-landmark supervision retains its advantage as recent history is removed, weakens coupling between predicted risk and EHR recording patterns, and improves reproducibility across retraining.
- **We introduce trajectory-aware evaluation for repeated prediction.** Separating predictive accuracy, temporal consistency, hazard variation, and trajectory direction reveals that improved agreement between successive predictions can coexist with smoother hazards and weak within-patient risk direction.
- **We identify trajectory dynamics as a distinct modeling target.** Repeated supervision improves accuracy and consistency without enforcing informative risk evolution. A ranking-augmented proof of principle on EHRSHOT shows that explicit cross-landmark objectives can alter trajectory direction, motivating dynamic models that learn relationships between predictions over time.

2 Related work

Dynamic survival prediction and landmarking. Dynamic survival prediction updates prognosis as new longitudinal information becomes available. Classical approaches include joint models and landmarking, which re-baselines survival prediction at selected times among patients still at risk (Houwelingen, 2007). Landmark supermodels and later machine-learning extensions share information across landmark-specific risk sets while allowing updated covariates (Tanner et al., 2021). We use the same re-baselining principle, but study landmarking as sparse supervision for shared

deep EHR models. We question the number and placement of supervised prediction times, and how it affects performance, robustness, and repeated prediction dynamics.

Deep longitudinal survival models and EHR representations. DeepHit models discrete event-time distributions with a ranking objective (Lee et al., 2018), while Dynamic-DeepHit extends this setting to longitudinal measurements using recurrent representations and auxiliary prediction objectives (Lee et al., 2020). Continuous-time methods instead model latent patient evolution directly, for example with controlled differential equations or neural trajectory models (Brouwer et al., 2019; Bleistein et al., 2024; Zeng et al., 2025). In parallel, EHR foundation models scale representation learning over long, irregular event streams (Li et al., 2025; Shmatko et al., 2025). MOTOR is particularly relevant because its pretraining objective is itself time-to-event based (Steinberg et al., 2023), and it already leverages dense supervision by using a collection of prediction times per patient. More broadly, these methods modify the representation, architecture, or objective, while multi-landmark supervision is complementary and remains model- and objective-agnostic.

Temporal consistency and trajectory-level behavior. Temporal-consistency methods explicitly relate predictions made at successive patient states. Maystre & Russo (2022) derive temporal-difference survival objectives, while later work improves end-to-end scalability and extends these constraints to irregular observation intervals (Vieyra & Frossard, 2024; Fah et al., 2026). We take a different perspective: rather than enforcing consistency, we ask whether it emerges from multi-landmark supervision and whether it reflects informative temporal behavior. This distinction matters because agreement alone can be achieved by flat predictions. We therefore evaluate temporal consistency jointly with hazard variation and patient-level trajectory direction.

EHR-specific robustness and evaluation. Structured EHRs are long, irregular, and shaped by the recording process itself. Prior work shows that context length, repeated codes, irregular temporal coverage, representation choice, and data provenance can materially affect model behavior (Wornow et al., 2025; Chen et al., 2026; Shmatko et al., 2025). Broader evaluations of EHR foundation models likewise argue that current evaluation strategies are insufficient to establish clinical utility (Wornow et al., 2023b). Survival models are still commonly evaluated with population-level concordance and Brier scores at fixed horizons, which do not characterize longitudinal behavior. Our evaluation therefore combines standard survival metrics with record truncation, sensitivity to recording density and redundancy, retraining reproducibility, temporal consistency, hazard variation, and patient-level risk trajectories.

3 METHODS

We formalize *sparse multi-landmark training* and define budget-matched controls to assess its effects. We also introduce an evaluation framework spanning prediction accuracy, robustness to EHR recording patterns, and within-patient risk dynamics. This framework tests whether better predictions at individual visits translate into more informative risk adjustments over time.

3.1 SPARSE LANDMARK SUPERVISION

Prediction from an observed history. On longitudinal electronic health records, a survival model is a function of a patient's history up to a prediction time. Patient p is represented by a chronologically ordered token stream $\mathcal{X}_p = \{(c_{p,i},\, t_{p,i})\}_{i=1}^{n_p}$, where $c_{p,i}$ is a coded clinical event and $t_{p,i}$ its timestamp. Let T_p^{ev} and T_p^{cens} denote event and censoring times, with $T_p^{\text{end}} = \min(T_p^{\text{ev}}, T_p^{\text{cens}})$ and $\delta_p = \mathbf{1}_{T_p^{\text{ev}} \le T_p^{\text{cens}}}$.

At landmark $t_{p,a}$, the model uses the available history $\mathcal{H}_p(t_{p,a})$, including demographics, to estimate

$$S_p^*(u \mid t_{p,a}) = Pr\big(T_p^{\text{ev}} > t_{p,a} + u \mid T_p^{\text{ev}} > t_{p,a}, \mathcal{H}_p(t_{p,a})\big)\,, \tag{1}$$

The model estimates Equation 1 with $S_{\theta,p,a}(u)$, using the observed label:

$$(y_{p,a}, \delta_{p,a}) = (T_p^{\text{end}} - t_{p,a}, \delta_p).$$

One-landmark training supervises only each patient's latest eligible history, whereas clinical deployment may require predictions at any visit. Sparse multi-landmark training extends supervision across

prediction times, changing both the observed history and the residual-time label while preserving the patient's outcome indicator. We refer to patients with an observed event ($\delta_p = 1$) as *cases* and those censored without an observed event ($\delta_p = 0$) as *controls*. At a given landmark and forecast horizon h, the *event* term includes observations with $y_{p,a} \leq h$ and $\delta_p = 1$, whereas the *at-risk* term includes observations with $y_{p,a} > h$. Observations censored at or before h have unknown event status at that horizon.

Sparsity. Predicting a residual time at every recorded event is the dense limit of landmark supervision. When all examples contribute to training, densely recorded patients receive more weight, and closely spaced events produce highly overlapping inputs. Instead, we sample a bounded pool of prediction times and select a sparse subset for training.

Eligibility. Each landmark coincides with an observed coded event. We require at least three codes up to event (including landmark event) and a minimum gap of three months ($g = 91.3$ days) before T_p^{end}. These fixed thresholds impose a minimum history length and limit access to documentation close to the event or censoring time. Eligibility is identical across training configurations, and patients without an eligible position are excluded.

Landmark sampling. We cap the pool at ten landmarks to bound its size, and build it using one of two stratified sampling policies. *Position-stratified* sampling partitions eligible event indices into contiguous bins of approximately equal size and samples one position uniformly from each bin. *Time-stratified* sampling partitions the eligible calendar interval into equal-width time bins, samples one target time uniformly per bin, and selects the nearest eligible event. Duplicate sampled indices are removed. The former sampling policy distributes supervision across recorded events, while the latter distributes it across elapsed time, thus limiting concentration within documentation bursts. We specify the policy used for each dataset in the experimental setup (Appendix A.2; Table 3). Across experiments, we use a fixed patient-specific seed for sampling.

Training configurations. Our baseline, one-landmark training (1L), supervises each patient at the latest eligible position. When we vary the number of landmarks L under a fixed optimizer budget, $L = 3$ captures most of the calibration gain, although additional landmarks could further improve discrimination (Table 6). We therefore adopt three-landmark training (3L) as a trade-off between sparsity and predictive performance.[1] The resulting trajectory coverage is reported in Table 3. During training, validation uses each patient's latest eligible landmark. At test time, we use the ten-landmark pool on held-out patients as a primary assessment of generalization across prediction times.

3.2 Model families

We assess sparse multi-landmark training across eight tasks from two health systems (Appendix A.1) and six model families. We select encoders with different inductive biases: a GRU models sequential dependencies through recurrence (Cho et al., 2014), a bidirectional Transformer captures content-dependent interactions through self-attention (Vaswani et al., 2017), and GRU-ODE explicitly models latent evolution between irregularly spaced events (Brouwer et al., 2019). This comparison tests whether the benefits of multi-landmark supervision extend across different approaches to modeling event order, interactions within the history, and elapsed time.

A linear probe on frozen MOTOR embeddings (Steinberg et al., 2023) further isolates supervision changes from encoder adaptation. These families share the same piecewise-exponential survival head as MOTOR. GRU-based DeepHit (Lee et al., 2018) and adapted Dynamic-DeepHit (Lee et al., 2020) provide complementary discrete-time configurations with ranking objectives, testing whether the landmark effect depends on the survival formulation. Appendix A.3-A.4 detail the architectural specifications and objectives.

[1] We partition the ordered ten-landmark pool into three contiguous bins and retain the midpoint position of each bin. Patients with fewer than three pooled positions contribute all available positions.

3.3 EVALUATION FRAMEWORK

We introduce an evaluation framework including prediction accuracy, robustness, and within-patient risk dynamics. Throughout our experiments, we use multi-landmarking evaluation (i.e., the ten-landmark test pool) as the primary setting. More details in Appendix B.

Prediction accuracy: Does multi-landmark training improve prediction? We measure overall discrimination with Harrell's concordance index (C-index) and discrimination up to each horizon $h \in \{1, 3, 5, 10\}$ years with Uno's concordance. Uno's estimator uses inverse probability of censoring weighting (IPCW) to adjust ranking comparisons for incomplete follow-up. Likewise, we use IPCW Brier scores to measure probabilistic prediction error at each horizon, summarized by their trapezoidal average (Integrated Brier Score, namely IBS). Censoring weights come from a shared Kaplan–Meier estimate fitted on the evaluation rows, ensuring that all models are scored using the same reference.

Robustness: Are predictions less sensitive to recording and retraining? We assess sensitivity to recording patterns, removal of recent history, and changes in the training sample. To quantify coupling to the recording process, we compute the age-adjusted partial Spearman correlation between predicted log cumulative hazard and two properties of the observed record, code *density* and code *redundancy*, among patients known to remain event-free through the forecast horizon. Lower absolute correlation indicates weaker coupling between predicted risk and how the record was documented. We quantify reproducibility using fold variance (Appendix C.5).

Risk dynamics: Are successive predictions consistent and informative?
Temporal consistency. Accuracy at individual visits does not establish whether successive predictions agree. Multi-landmark evaluation allows us to examine this by following predictions across each patient's history. While prior work uses temporal consistency as a training constraint (Maystre & Russo, 2022), here, we ask whether multi-landmark training improves consistency without explicitly enforcing it. Successive predictions describe overlapping future periods from different starting times. To compare them, we condition the earlier curve on remaining event-free until the later landmark. For an earlier prediction S_i and landmark interval Δ, the transported curve is $\widetilde{S}_i(u) = \frac{S_i(u+\Delta)}{S_i(\Delta)}$. We summarize disagreement between the transported and directly predicted survival curves using total-variation distance, m_{TV} (Appendix C), with lower values indicating greater temporal consistency.

Direction of risk. We regress each patient's predicted risk $r_{p,a}(h) = 1 - S_{p,a}(h)$ on landmark calendar time, where h is a fixed forecast horizon measured from each landmark. Within each training fold, we report median slopes $\widehat{b}_p(h)$ separately for observed cases and censored controls, limiting the influence of extreme slope estimates from early-trajectory landmarks. Positive case slopes indicate rising risk as the observed event approaches. We measure case–control separation using individual patient slopes:

$$\mathrm{slopeAUC}(h) = \Pr\Big(\widehat{b}_{\mathrm{case}}(h) > \widehat{b}_{\mathrm{control}}(h)\Big) + \frac{1}{2}\Pr\Big(\widehat{b}_{\mathrm{case}}(h) = \widehat{b}_{\mathrm{control}}(h)\Big). \tag{2}$$

Together, these measures characterize how risk evolves within patients and whether that evolution differs by outcome: median slopes describe the direction of change, while slope AUC quantifies the dynamics differentiation (Appendix C.4).

Across all evaluation axes, we compute metrics separately within each fold and report their mean and standard deviation over five training folds. Comparisons are paired by fold, with counts indicating how consistently the effect direction holds.

4 RESULTS

4.1 MULTI-LANDMARK SUPERVISION IMPROVES PREDICTION ACROSS THE PATIENT TRAJECTORY

Three-landmark training improves prediction across tasks and model families. We first evaluate whether supervising each patient at three landmarks rather than one improves prediction across

Table 1: **Prediction gain from three-landmark training under multi-landmark evaluation.** Paired difference between three-landmark (3L) and one-landmark (1L) training for Harrell's C-index and integrated Brier score (IBS), in percentage points, evaluated on the ten-landmark test pool. Values report the mean difference across five training folds ± standard deviation. Positive values indicate improved performance with 3L; the IBS difference is oriented so that a reduction in prediction error is positive. Absolute 1L and 3L performance is reported in Table 5.

	EHRSHOT			MIMIC-IV				
Model	AMI	HPT	PC	BC	ADRD	DB	HPT	LC
(a) Harrell's C								
GRU	−1.11 ±3.36	3.46 ±3.95	4.29 ±2.81	2.59 ±6.75	6.10 ±4.21	3.50 ±0.45	1.55 ±0.35	9.95 ±15.42
Transformer	0.19 ±4.41	0.68 ±5.98	5.90 ±2.23	3.87 ±5.59	2.65 ±1.67	3.22 ±2.17	2.01 ±0.40	5.00 ±2.76
GRU-ODE	5.19 ±4.80	11.96 ±4.22	6.37 ±10.22	8.17 ±1.63	10.50 ±1.98	13.37 ±1.55	6.09 ±3.06	2.31 ±1.36
DeepHit	0.48 ±1.21	5.81 ±1.06	10.81 ±2.97	3.05 ±3.49	5.64 ±2.64	2.15 ±1.08	0.91 ±0.69	2.99 ±0.97
Dynamic-DeepHit	0.76 ±3.83	3.97 ±2.11	10.17 ±3.65	4.53 ±4.04	5.94 ±3.40	1.45 ±0.63	0.63 ±1.08	4.70 ±1.03
MOTOR-Probe	16.46 ±0.63	7.76 ±1.48	3.30 ±0.70	1.71 ±0.80	1.33 ±0.24	−0.34 ±0.16	−0.35 ±0.09	1.70 ±0.32
(b) Integrated Brier score (IBS)								
GRU	0.28 ±1.75	13.30 ±7.65	4.82 ±1.51	14.14 ±29.78	3.22 ±1.69	1.83 ±0.52	1.86 ±0.52	17.77 ±30.73
Transformer	6.34 ±6.31	12.37 ±18.42	0.65 ±2.82	2.70 ±2.93	7.49 ±9.20	1.71 ±2.00	2.75 ±1.01	0.70 ±1.46
GRU-ODE	5.90 ±4.74	5.91 ±3.94	0.79 ±6.60	9.77 ±8.19	3.85 ±0.85	2.44 ±0.54	4.86 ±1.57	2.02 ±0.88
DeepHit	6.57 ±7.33	16.81 ±7.11	12.61 ±1.85	13.32 ±19.20	1.82 ±0.56	1.13 ±0.69	1.34 ±0.52	12.98 ±6.03
Dynamic-DeepHit	10.39 ±6.13	21.83 ±5.14	12.33 ±2.15	12.76 ±19.58	1.89 ±0.53	0.72 ±0.56	1.15 ±0.45	10.01 ±6.78
MOTOR-Probe	5.78 ±0.94	11.20 ±1.41	11.46 ±1.46	0.41 ±0.08	0.59 ±0.08	0.30 ±0.05	0.79 ±0.02	0.47 ±0.07

Tasks: AMI = Acute myocardial infarction, HPT = Hypertension, PC = Pancreatic cancer, BC = Breast cancer, ADRD = Alzheimer's and relative dementia, DB = Diabetes, LC = Lung cancer.

the patient trajectory. On the ten-landmark test pool, three-landmark training (3L) improves the integrated Brier score in all 48 task–model configurations and Harrell's C-index in 45 of 48, with median gains of 4.34 and 3.48 percentage points, respectively (Table 1). The improvement spans both health systems and all six model families, including models based on recurrence, self-attention, continuous-time dynamics, discrete-time survival objectives, and frozen pretrained representations. Multi-landmark supervision therefore improves predictive performance across substantially different model formulations rather than benefiting a single architecture or survival objective. Per-task and per-model breakdowns are given in Table 5.

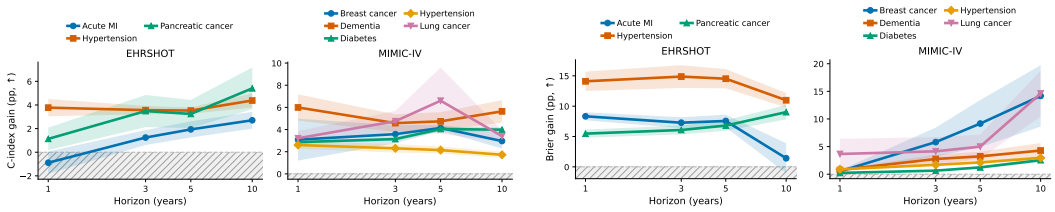

Figure 1: **Prediction gains persist across forecast horizons.** Difference between 3L and 1L for horizon-specific Uno's concordance (left) and Brier score (right). Positive values indicate better performance with 3L. Results are shown at 1-, 3-, 5-, and 10-year forecast horizons; the number of event and at-risk landmarks available at each horizon is reported in Table 7.

The improvement persists from short- to long-term prediction. The benefit of 3L extends across the 1-, 3-, 5-, and 10-year forecast horizons (Figure 1). Its magnitude varies by health system and metric: on MIMIC-IV, the Brier-score gain becomes larger toward longer horizons while the discrimination gain remains comparatively stable; on EHRSHOT, the discrimination gain tends to increase with the horizon while the Brier-score improvement is already substantial at shorter horizons. Multi-landmark supervision therefore retains its predictive advantage for long-range risk estimation, rather than improving only near-term forecasts. This regime is most relevant to conditions or applications that require identifying elevated risk well before an event, such as screening and preventive intervention (see task-decomposition in Figures 7-8).

4.2 The gain persists under controlled budgets and EHR perturbations

The gain is not explained by additional optimization or input tokens. 3L exposes each patient at several prediction times, but its benefit does not require a larger training budget. We first fix the

number of optimizer updates for the MOTOR-Probe, so that 1L and 3L train on the same patients for the same number of updates. Under this constraint, 3L improves IBS on all eight tasks and Harrell's C on seven out of eight (Table 9).

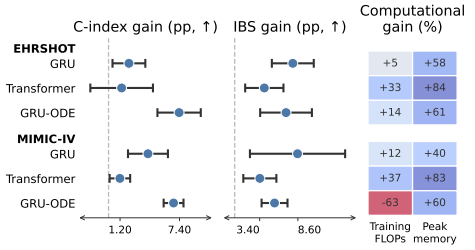

Figure 2: **3L improves prediction under a matched input-token budget.** We compare 3 shorter history contexts with 1 context 3× as long, holding input-token budget and model parameter count fixed. Positive values indicate 3L is better. Shorter contexts also allows for fewer FLOPs and peak memory per sample.

We obtain the same qualitative conclusion when matching the input-token budget: distributing the same number of tokens across three shorter contexts improves IBS across all evaluated task–encoder configurations and discrimination in all of them (Figure 2; Table 8). Shorter contexts reduce FLOPs and peak memory per training example for all architectures. Still, some GRU-ODE configurations require more optimization steps to converge under 3L, explaining the higher training FLOPs values for this model. Multi-landmark supervision therefore improves predictive performance at a fixed input budget while reducing per-example computation.

Multi-landmark supervision can partly compensate for fewer training patients. Next, we train MOTOR-Probe 3L with fewer patients and compare it with 1L trained on the full cohort. On all three EHRSHOT tasks, 3L trained on only half of the patients still matches or exceeds the full-cohort 1L baseline (Figure 9). This trade-off is specific to EHRSHOT in our experiments, as substantially shorter trajectories from MIMIC-IV provide less additional temporal supervision per patient. This shows that additional temporal supervision can, in settings with informative longitudinal histories, partly compensate for fewer training patients in small cohorts.

The advantage persists when recent clinical history is removed. Then, we ask whether the gain depends on documentation recorded close to the outcome (Appendix D). Without retraining, we progressively increase the minimum interval between the last code available to the model and the event or censoring time. If the advantage of multi-landmark supervision were driven primarily by outcome-proximal information, it should diminish as we remove that information. Instead, the IBS advantage is maintained and generally becomes larger as the interval increases, while the discrimination advantage remains positive and comparatively stable (Figure 3). Multi-landmark supervision therefore retains its advantage when predictions must be made from more distant patient history, arguing against proximity to the recorded endpoint as the explanation for the gain. Figures 10-11 further break down these gains per model and task.

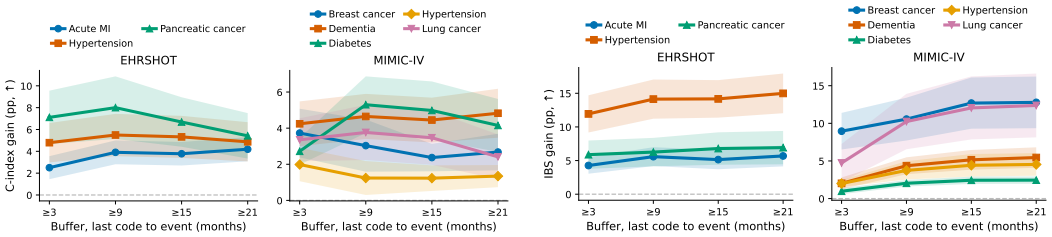

Figure 3: **Prediction gains persist as recent patient history is withheld.** Difference between 3L and 1L training in concordance (left) and IBS (right) as the minimum interval between the last observed code and the event is increased at inference. Positive values indicate better performance with 3L.

Multi-landmark supervision shrinks coupling to EHR recording patterns. EHRs encode not only patient state but also how care is delivered and documented. Therefore, we measure whether predicted risk becomes less associated with two properties of the observed record: code density and code redundancy. Among patients known to remain event-free through the forecast horizon, 3L generally reduces the magnitude of the statistical coupling to the recording process, with the clearest reductions on EHRSHOT (Figure 4; Table 10). The effect is especially consistent for code redundancy, suggesting that multi-landmark supervision reduces reliance on repeated documentation patterns that can confound disease severity.

Predictions are also more reproducible across retraining. Finally, three-landmark training yields more stable predictions across retraining, with lower fold-to-fold prediction variability in 26 of 31 task–model configurations at the 5-year horizon (Appendix 11). Multi-landmark supervision therefore improves average predictive performance while also increasing robustness to record perturbations and stability across retraining.

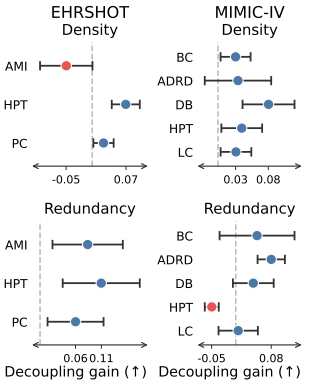

Figure 4: **Coupling with EHR recording patterns.** Positive values means lower coupling with 3L; dashed line denotes no change.

4.3 Multi-landmark supervision improves temporal consistency, motivating trajectory-aware evaluation

Standard survival metrics evaluate predictions at individual prediction times, but do not characterize how repeated predictions for the same patient evolve. This distinction matters when models are queried throughout follow-up: successive predictions should not only be accurate in isolation, but should remain coherent as the observed history changes and preserve clinically meaningful variation over time. We therefore ask whether multi-landmark supervision improves agreement between successive predictions, whether this agreement preserves variation in the predicted hazard, and whether patient-level risk trajectories differ according to outcome.

3L makes successive predictions more consistent. We first transport each prediction to the time of the patient's next landmark and compare it with the newly generated prediction (Section 3.3). 3L reduces this disagreement in 28 of 32 task–model combinations, including all 12 combinations on EHRSHOT (Table 12). The effect extends across model families and both health systems, showing that supervision distributed across prediction times produces substantially more consistent sequences of predictions.

Improved consistency is accompanied by smoother predicted hazards. Consistency alone does not establish that a model remains responsive to changes in patient history. A prediction that varies little over time is easier to transport consistently, even if that stability reflects reduced temporal responsiveness. We therefore measure hazard dispersion alongside consistency. 3L reduces hazard dispersion in 27 of 32 combinations, and 25 combinations show both greater consistency and smoother hazards (Table 12). The two changes also track one another across configurations (Figure 5). We interpret this as a co-occurrence rather than a causal relationship: improved agreement is frequently accompanied by flattening, so temporal consistency alone cannot establish that repeated predictions capture more informative dynamics.

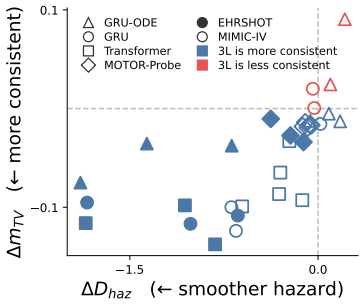

Figure 5: **Temporal consistency vs. predicted hazards variation.**

Trajectory direction provides a complementary test of temporal behaviour. To distinguish stability from informative change, we directly measure how predicted risk evolves within each patient (Section 3.3). At a fixed forecast horizon, a positive case slope corresponds to increasing predicted risk as the observed event approaches, while slope AUC (Equation 2) measures whether case trajectories tend to increase more than control trajectories. Under 1L, the median case slope is negative in 30 of 31 task–model combinations; under 3L, it remains negative in 24 of 31 (Table 13). Case–control slope separation also remains modest, with slope AUC concentrated near 0.5. Thus, multi-landmark supervision substantially improves agreement between successive predictions, while direct trajectory measurements reveal temporal behaviour that prediction accuracy and consistency do not capture.

5 Discussion

Sparse multi-landmark supervision addresses a simple mismatch in EHR survival modeling: models are commonly trained at one prediction time per patient but deployed repeatedly throughout follow-

up. Our results show that broadening supervision across prediction times improves predictive performance, robustness to record perturbations, reproducibility across retraining, and agreement between successive predictions. Multi-landmark supervision reveals which temporal properties emerge from repeated supervision and which require explicit cross-landmark constraints.

Repeated supervision does not imply relational supervision. Multi-landmark training exposes the model to different states of the same patient, but each landmark remains an independent supervision target. Shared parameters are enough to improve accuracy and temporal consistency, yet consistency often increases alongside smoother hazards. Thus, better agreement does not imply more informative risk trajectories.

Trajectory-aware evaluation is therefore necessary for repeatedly deployed models. Standard discrimination and Brier scores assess predictions at individual times, while consistency measures agreement across landmarks. Neither alone captures whether risk evolves meaningfully within a patient. Accuracy, consistency, and trajectory direction are therefore complementary properties, to measure performance as new information arrives. More generally, any survival model repeatedly applied to an evolving entity has the same evaluation problem.

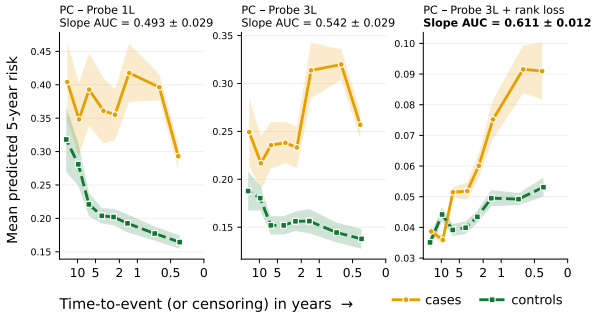

Figure 6: **Within-patient risk trajectories from MOTOR-Probe.** Predicted 5-year risk for pancreatic-cancer cases and controls as the event approaches. The ranking loss increases slope, and slope AUC.

Explicit within-patient ordering. The standard survival objective does not constrain how a patient's predictions should be ordered across landmarks. We therefore add a within-patient ordering objective to three-landmark MOTOR-Probe training (Appendix H). This increases case–control slope separation across all three EHRSHOT tasks (Table 12); for pancreatic cancer, case risk rises toward the event while controls remain comparatively stable (Figure 6). We treat this as a proof of principle that explicitly coupling predictions across time can alter trajectory direction.

Temporal structure requires explicit modeling. These findings distinguish training *at* multiple times from training *across* them: the former broadens supervision, while the latter constrains temporal relationships. Future work should combine explicit within-patient objectives with architectures that model continuous latent evolution, for example by coupling trajectory-level constraints to neural ODE-style latent dynamics rather than only to the output risk. Such objectives must also avoid trivial solutions: consistency losses can be satisfied by nearly flat hazards, so temporal regularization should preserve responsiveness as well as agreement. Landmark placement is another design variable, since bursty EHRs can yield many prediction times but little elapsed time over which dynamics can be learned. Finally, evaluation should move beyond linear slopes toward trajectory-level estimands defined directly on irregular prediction sequences.

Overall, sparse multi-landmark supervision improves EHR survival prediction, but better predictions at multiple snapshots do not imply learned temporal relationships. Explicitly modeling and evaluating those relationships is a key next step toward more informative longitudinal risk prediction.

AI USE STATEMENT

Large language models and coding agents were used to assist with code generation, manuscript structuring, initial drafting, and LaTeX preparation. The study design, scientific hypotheses, experiments, analyses, literature review, interpretation of results, and verification of all reported findings were performed by the authors, who take full responsibility for the content of the paper.

Ethics statement

This study uses de-identified electronic health record data from EHRSHOT and MIMIC-IV under their respective data-access and usage requirements. The work is retrospective and methodological and does not involve prospective clinical intervention.

Reproducibility statement

We provide detailed descriptions of cohort construction, landmark sampling, model architectures and objectives, hyperparameter search spaces, training procedures, evaluation metrics, and experimental controls in the supplementary material. All experiments use publicly available datasets, MIMIC-IV (Johnson et al., 2024) and EHRSHOT (Wornow et al., 2023a), subject to their respective access requirements. Random seeds and the definitions of the trajectory, robustness, and recording-process metrics are also documented in the appendix. We intend to release the complete implementation, including preprocessing, multi-landmark training, evaluation, and experiment scripts, in a public Github repository following publication.

References

Linus Bleistein, Van Tuan Nguyen, Adeline Fermanian, and Agathe Guilloux. Dynamic Survival Analysis with Controlled Latent States. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 4160–4204. PMLR, 2024. URL https://proceedings.mlr.press/v235/bleistein24a.html.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. *GRU-ODE-Bayes: continuous modeling of sporadically-observed time series*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Tianyi Chen, Mingcheng Zhu, Zhiyao Luo, and Tingting Zhu. Cross-Representation Benchmarking in Time-Series Electronic Health Records for Clinical Outcome Prediction. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7076–7080, 2026. doi: 10.1109/ICASSP55912.2026.11462127.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179.

Harrison Fah, Russell Greiner, and Roger A. Dixon. Temporally consistent survival prediction for non-uniform longitudinal data. *Journal of Biomedical Informatics*, 176:104996, 2026. doi: 10.1016/j.jbi.2026.104996.

Charles Gadd, Krishna Gokhale, Aditya Acharya, Jennifer Cooper, Francesca Crowe, Leah Fitzsimmons, Thomas Jackson, Krishnarajah Nirantharakumar, and Christopher Yau. SurvivEHR: a competing risks, time-to-event foundation model for multiple long-term conditions from primary care electronic health records. *npj Digital Medicine*, 9(1):546, 2026. doi: 10.1038/s41746-026-02709-z.

Hans C. Van Houwelingen. Dynamic Prediction by Landmarking in Event History Analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007. URL https://ideas.repec.org//a/bla/scjsta/v34y2007i1p70-85.html.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, 2024. doi: 10.13026/kpb9-mt58. Version 3.1.

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), 2018. doi: 10.1609/aaai.v32i1.11842.

Changhee Lee, Jinsung Yoon, and Mihaela van der Schaar. Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2020. doi: 10.1109/TBME.2019.2909027.

Hao Li, Bowen Deng, Chang Xu, ZhiYuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. MIRA: Medical Time Series Foundation Model for Real-World Health Data. 2025. URL https://openreview.net/forum?id=Auy2DmlJBO.

Lucas Maystre and Daniel Russo. Temporally-Consistent Survival Analysis. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10671–10683. Curran Associates, Inc., 2022. doi: 10.52202/068431-0775.

Davide Placido, Bo Yuan, Jessica X. Hjaltelin, Chunlei Zheng, Amalie D. Haue, Piotr J. Chmura, Chen Yuan, Jihye Kim, Renato Umeton, Gregory Antell, Alexander Chowdhury, Alexandra Franz, Lauren Brais, Elizabeth Andrews, Debora S. Marks, Aviv Regev, Siamack Ayandeh, Mary T. Brophy, Nhan V. Do, Peter Kraft, Brian M. Wolpin, Michael H. Rosenthal, Nathanael R. Fillmore, Søren Brunak, and Chris Sander. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature Medicine*, 29(5):1113–1122, 2023. doi: 10.1038/s41591-023-02332-5.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, 2025. doi: 10.1038/s41586-025-09529-3.

Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. MOTOR: A Time-to-Event Foundation Model For Structured Medical Records. 2023. URL https://openreview.net/forum?id=NialiwI2V6.

Kamaryn T. Tanner, Linda D. Sharples, Rhian M. Daniel, and Ruth H. Keogh. Dynamic Survival Prediction Combining Landmarking with a Machine Learning Ensemble: Methodology and Empirical Comparison. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184 (1):3–30, 2021. doi: 10.1111/rssa.12611.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Mariana Vargas Vieyra and Pascal Frossard. Deep End-to-End Survival Analysis with Temporal Consistency, 2024. URL https://arxiv.org/abs/2410.06786.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d42db1f74df54cb992b3956eb7f15a6f-Abstract-Datasets_and_Benchmarks.html.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, 2023b. doi: 10.1038/s41746-023-00879-8.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Re, Sanmi Koyejo, and Nigam H. Shah. Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHRs, 2025.

Sihang Zeng, Lucas Jing Liu, Jun Wen, Meliha Yetisgen, Ruth Etzioni, and Gang Luo. TrajSurv: Learning Continuous Latent Trajectories from Electronic Health Records for Trustworthy Survival Prediction. In *Proceedings of the 10th Machine Learning for Healthcare Conference*. PMLR, 2025. URL https://proceedings.mlr.press/v298/zeng25a.html.

Andrew Zhang, Tong Ding, Sophia J. Wagner, Caiwei Tian, Ming Y. Lu, Rowland Pettit, Joshua E. Lewis, Alexandre Misrahi, Dandan Mo, Long Phi Le, and Faisal Mahmood. A multimodal and temporal foundation model for virtual patient representations at healthcare system scale, 2026.

A Supplementary Methods

A.1 Datasets and cohort construction

EHRSHOT. EHRSHOT contains 6,739 de-identified Stanford Medicine patients and provides canonical patient-level training, validation, and test splits (Wornow et al., 2023a). We re-derived three benchmark phenotypes as right-censored time-to-event tasks: acute myocardial infarction (AMI), essential hypertension (HPT), and pancreatic cancer (PC). Event definitions comprise the corresponding task concept and its descendants in the OMOP hierarchy. Event time is the first qualifying occurrence in the available record; patients without a qualifying event are censored at the end of observed follow-up.

Inputs comprise timestamped conditions, drug exposures, procedures, and measurements. Concepts observed in at least ten eligible patients of the canonical training split are retained; all remaining concepts map to an unknown token. After requiring an available token history and at least one eligible landmark, the AMI, HPT, and PC cohorts contain 3,640, 2,163, and 3,656 patients, respectively. Canonical test patients remain fixed. Training and validation patients are pooled and repartitioned into five event-stratified development folds.

MIMIC-IV. We use MIMIC-IV v3.1 converted to OMOP CDM v5.4 (Johnson et al., 2024), retaining conditions, drug exposures, procedures, and laboratory measurements. Cases are identified from raw ICD diagnosis codes using task-specific prefix definitions, with event time given by the first qualifying diagnosis. We require prior observed history before the event and apply task-specific phenotype filters, including exclusion of type-1 diabetes from the diabetes cohort and restriction of breast cancer to female patients. Controls have no qualifying diagnosis for the corresponding phenotype and satisfy the same minimum observation requirements.

Controls are selected by propensity-score matching before train–validation–test splitting. Propensity scores are estimated with logistic regression after median imputation and standardization of numeric variables and one-hot encoding of categorical variables. Matching uses nearest neighbors on the scalar propensity score without a caliper or common-support trimming. Covariates include combinations of age, sex, race, observed follow-up duration, and mortality status, with the exact set varying slightly by task. The resulting cohort is then split at the patient level using event-, age-, and sex-stratified sampling.

For model inputs, OMOP concept identifiers from the four retained domains are converted to domain-specific tokens. The vocabulary is constructed using training patients only and retains concepts observed in at least ten patients; all remaining concepts map to an unknown token. Rows with missing patient identifiers, timestamps, or concept identifiers are discarded, and events occurring after the prediction cutoff are excluded.

Input representation and demographics. For both health systems, longitudinal inputs are chronologically ordered clinical concept identifiers with timestamps. Seven demographic variables are evaluated at each landmark: sex, five race indicators, and age at the landmark. Age therefore changes across prediction times for the same patient. Cohort-level statistics and landmark coverage are reported in Tables 2 and 3.

A.2 Landmark construction

Eligibility. For patient p, let $t_1 \leq \cdots \leq t_{n_p}$ denote recorded event times and T_p^{end} the observed event or censoring time. A landmark must coincide with an observed event, contain at least three codes in the available history, and precede T_p^{end} by at least $g = 91.3$ days. These criteria are identical across landmark-training configurations. Patients without an eligible landmark are excluded.

Landmark pools. For each patient, we construct a deterministic pool of at most ten eligible prediction times using a fixed patient-specific seed. Two sampling policies are used because the two datasets have different temporal geometry.

Table 2: **Cohort and longitudinal-record characteristics by task.** Statistics are reported for all patients with at least one eligible landmark across the training, validation, and test splits. Each trajectory includes all recorded codes up to and including the patient's latest eligible landmark. Record span is measured from the first to the last dated code; code density is the number of codes per month of observed span, and redundancy is $1 - \#\text{unique}/\#\text{codes}$. Values summarize the patient-level distributions within each task, with trajectory length, span, density, and redundancy reported as medians.

					Median trajectory		Codes / month		Redundancy (%)	
Dataset	Task	Patients	Cases / controls	Prevalence (%)	Codes	Span (y)	Cases	Controls	Cases	Controls
EHRSHOT	AMI	3,640	375 / 3,265	10.3	5,545	7.8	64.8	69.3	94.4	96.1
	HPT	2,163	526 / 1,637	24.3	3,074	6.0	43.4	58.9	92.5	95.0
	PC	3,656	191 / 3,465	5.2	5,748	8.0	54.5	71.8	93.2	96.1
MIMIC-IV	BC	6,413	723 / 5,690	11.3	239	2.0	122.6	25.9	47.3	60.3
	ADRD	9,561	2,274 / 7,287	23.8	296	1.6	74.7	27.4	57.3	65.0
	DB	16,559	2,438 / 14,121	14.7	227	1.7	171.8	24.3	51.7	59.1
	HPT	14,203	4,778 / 9,425	33.6	139	0.1	456.6	57.3	42.3	50.0
	LC	8,822	1,155 / 7,667	13.1	264	1.1	186.4	43.8	58.8	62.3

Tasks: AMI = Acute myocardial infarction, HPT = Hypertension, PC = Pancreatic cancer, BC = Breast cancer, ADRD = Alzheimer's disease and related dementia, DB = Diabetes, LC = Lung cancer.

Table 3: **Landmark-set size and trajectory coverage.** For each task, we report the number of training landmarks under one-landmark (1L) and three-landmark (3L) supervision, the resulting number of patient–landmark samples, and the portion of each longitudinal record covered by the union of 512-code context windows ending at those landmarks. Training and validation sample counts are averaged over five folds; test sets are shared across folds. The ten-landmark test pool uses approximately ten prediction times per patient, distributed over calendar time on EHRSHOT and by event index on MIMIC-IV. Coverage is summarized across training patients as the median fraction of observed codes, years of record, and fraction of elapsed trajectory time represented by the landmark windows, averaged over folds. Because 1L uses the latest eligible landmark, codes occurring after the latest 3L landmark may not be covered, so 3L does not necessarily maximize raw code coverage.

				Samples				Coverage, 512-code window		
Dataset	Task	Landmarks	Landmarks / patient	Train	Val	Test 1a	Test 10a	Codes (%)	Years	Span (%)
EHRSHOT	AMI	1a	1.00	1,970	492	1,178	9,224	9.2	0.93	18.7
		3a	2.98	5,865	1,466	1,178	9,224	13.3	4.49	69.9
	HPT	1a	1.00	1,167	292	704	5,341	16.4	0.95	29.7
		3a	2.97	3,462	866	704	5,341	21.0	3.89	77.2
	PC	1a	1.00	1,974	493	1,189	9,371	9.0	0.95	18.9
		3a	2.98	5,879	1,470	1,189	9,371	13.0	4.52	69.1
MIMIC-IV	BC	1a	1.00	4,627	1,157	629	6,163	100.0	0.62	100.0
		3a	2.98	13,812	3,453	629	6,163	84.8	1.15	97.1
	ADRD	1a	1.00	6,883	1,721	957	9,444	100.0	0.51	100.0
		3a	2.99	20,558	5,139	957	9,444	84.7	0.94	95.5
	DB	1a	1.00	11,922	2,980	1,657	16,188	100.0	0.55	100.0
		3a	2.98	35,524	8,881	1,657	16,188	84.9	0.95	96.5
	HPT	1a	1.00	10,237	2,559	1,407	13,591	100.0	0.06	100.0
		3a	2.97	30,388	7,597	1,407	13,591	85.5	0.05	93.1
	LC	1a	1.00	6,350	1,587	885	8,742	100.0	0.28	100.0
		3a	2.99	18,966	4,741	885	8,742	84.7	0.55	94.0

Tasks: AMI = Acute myocardial infarction, HPT = Hypertension, PC = Pancreatic cancer, BC = Breast cancer, ADRD = Alzheimer's disease and related dementia, DB = Diabetes, LC = Lung cancer.

For MIMIC-IV, *position-stratified sampling* partitions eligible event indices into approximately equal contiguous bins and samples one event from each bin. This distributes landmarks over the recorded event sequence.

For EHRSHOT, *time-stratified sampling* partitions the eligible calendar-time interval into equal-width bins, samples one target time within each bin, and maps it to the closest eligible recorded event. Duplicate event indices are removed. This distributes landmarks over elapsed time and limits concentration within dense documentation bursts.

One- and three-landmark training. One-landmark training (1L) uses the latest eligible prediction time for each patient. Three-landmark training (3L) is derived from the ordered ten-landmark pool by partitioning it into three contiguous groups and retaining the central position from each

group. Patients with fewer than three pooled landmarks contribute all available positions. At test time, the complete ten-landmark pool is used for the primary multi-landmark evaluation.

Input windows. At landmark index a, token-based models receive the most recent L events ending at that landmark, left-padded when fewer than L events are available. Continuous inter-event time is represented as $\log(1 + \Delta)$, where Δ is measured in years. Padding is excluded by a binary mask. Training-time token dropout replaces a real clinical concept with the unknown token with probability 0.1.

A.3 Model families and survival objectives

We evaluate six model families spanning recurrent, attention-based, continuous-time, discrete-time survival, and pretrained EHR representations. The GRU, Transformer, and GRU–ODE encoders use the same piecewise-exponential survival head, while DeepHit and Dynamic-DeepHit use a discrete event-time objective. MOTOR is evaluated through a frozen representation with a learned survival probe.

Piecewise-exponential survival head. The shared continuous-time head predicts log hazards on six intervals,

$$[0, 0.5),\ [0.5, 1),\ [1, 2),\ [2, 5),\ [5, 10),\ [10, \infty)$$

years. Horizon-specific survival and risk are obtained from the corresponding cumulative hazard. The scalar risk score used for overall Harrell concordance is the mean predicted log hazard across intervals.

GRU and Transformer. The GRU processes event embeddings and continuous inter-event times sequentially and pools the resulting hidden states with masked attention (Cho et al., 2014). The Transformer applies bidirectional self-attention only within the observed history ending at the landmark; no information after the prediction time is available to the encoder (Vaswani et al., 2017).

GRU–ODE. GRU–ODE explicitly separates latent evolution between observations from updates at recorded events (Brouwer et al., 2019). A neural ODE transports the hidden state over each observed time gap, after which a GRU cell incorporates the new event. Fixed-step fourth-order Runge–Kutta integration is used. Where selected during tuning, an auxiliary reconstruction objective encourages the transported state to anticipate the post-observation representation.

Frozen MOTOR probe. We train a linear survival probe on frozen 768-dimensional MOTOR representations (Steinberg et al., 2023). The landmark embedding is concatenated with demographics and mapped to the same six piecewise-exponential hazards used by the from-scratch encoders. Because the MOTOR encoder remains frozen, this comparison isolates the effect of downstream landmark supervision from encoder adaptation.

DeepHit. DeepHit predicts a discrete event-time distribution and combines a likelihood term with its pairwise ranking objective (Lee et al., 2018). For comparable patients (i, j) satisfying $y_i < y_j$ and $\delta_i = 1$, the ranking component encourages greater cumulative incidence for patient i at y_i:

$$\mathcal{L}_{\mathrm{DH}} = \mathcal{L}_{\mathrm{NLL}} + w_{\mathrm{rank}}\mathcal{L}_{\mathrm{rank}}, \tag{3}$$

with the standard exponential pairwise ranking formulation.

Adapted Dynamic-DeepHit. Dynamic-DeepHit extends DeepHit with a longitudinal recurrent representation and auxiliary prediction of subsequent measurements (Lee et al., 2020). We adapt the latter to structured EHR event streams through next-token prediction. Because all tasks contain a single event of interest, the original competing-risk output reduces to a single event-time head. The same 1L and 3L landmark construction is applied to DeepHit and Dynamic-DeepHit.

A.4 Optimization and model selection

Token encoders are optimized with AdamW, cosine learning-rate annealing, gradient clipping, early stopping, and a maximum of 200 epochs. Architecture-specific hyperparameters are selected with

Optuna over learning rate, representation width, depth, batch size, context length, dropout and pooling choices; GRU–ODE additionally searches the ODE width and reconstruction-loss weight.

For the from-scratch models, hyperparameters are selected on the one-landmark configuration and then inherited by the corresponding three-landmark baseline. This keeps architecture and optimization choices fixed across the primary 1L–3L comparison. The selected configuration is subsequently retrained across five development folds. MOTOR-Probe uses a fixed optimization configuration across runs.

DeepHit and Dynamic-DeepHit retain the selected recurrent encoder and additionally tune their survival-objective parameters. The same selected hyperparameters are used when comparing 1L and 3L within each model family.

All reported training runs used a single NVIDIA P100 GPU per model unless otherwise noted; the five fold-models were trained independently.

A.5 Hyperparameter search and model selection

For the GRU, Transformer, and GRU–ODE encoders, we use Optuna with a tree-structured Parzen estimator to search architecture and optimization hyperparameters. Trials are ranked by the mean validation Harrell concordance over the final ten training epochs, rather than the single best epoch, to reduce selection of transient performance spikes. Final configurations are then retrained across five development folds. Unless varied by the search, models are trained with AdamW, weight decay 0.01, gradient clipping at 1.0, cosine learning-rate annealing to 10^{-6}, and a maximum of 200 epochs. The same search spaces are used across EHRSHOT and MIMIC-IV; the number of trials varies by task and architecture. For MIMIC-IV, model hyperparameters are selected using one-landmark training.

Table 4 summarizes the searched parameters. Parameters not listed in the table are held fixed within a model family.

For DeepHit, the number of event-time bins is fixed at 16, the ranking temperature is fixed at $\sigma = 0.1$, and the prediction head is linear. Dynamic-DeepHit uses the same discrete-time survival head and additionally tunes the weight β of the next-token auxiliary objective. The MOTOR-Probe is a linear survival head on frozen MOTOR representations and uses a fixed configuration in the experiments reported in the main paper.

B Experimental Design and Budget Controls

The primary comparison changes the number of supervised prediction times while keeping the patient cohort and model family fixed. We use three complementary controls to separate the effect of landmark supervision from additional optimization or input exposure.

B.1 Primary evaluation population

The primary test block is the ten-landmark pool of held-out patients used throughout the main paper. All compared configurations are aligned on patient and landmark identifiers and scored on common valid rows. This multi-landmark evaluation reflects repeated deployment across the observed trajectory. A latest-landmark evaluation is retained as a supplementary comparison but is not used as the primary evaluation in the main text.

B.2 Matched optimizer budget

For the MOTOR-Probe control, 1L and 3L use the same patients, batches, and fixed number of optimizer updates. One landmark per patient is drawn at each epoch: 1L always supplies the same latest landmark, whereas 3L samples among the patient's three training landmarks. Early stopping is disabled for this comparison. The two settings therefore differ in the diversity of supervised prediction times, not in optimizer steps.

Table 4: **Hyperparameter search spaces.** Search spaces used for the model families evaluated in the paper. DeepHit and Dynamic-DeepHit retain the selected recurrent backbone and tune only the survival-head parameters shown below. MOTOR-Probe uses one fixed configuration in the main experiments; the probe search space implemented during development is reported for completeness.

Model	Parameter	Search space
GRU	Learning rate	$\{5\times10^{-6},\ 10^{-5},\ 5\times10^{-5},\ 10^{-4},\ 3\times10^{-4},\ 10^{-3},\ 3\times10^{-3}\}$
	Embedding dimension	$\{32, 64, 128, 256\}$
	Hidden dimension	$\{32, 64, 128, 256, 512\}$
	Number of layers	$\{1, 2, 3\}$
	Batch size	$\{32, 64, 128, 256\}$
	Context length	$\{64, 256, 512\}$
Transformer	$(d_{\text{model}}, n_{\text{head}})$	$\{(64, 2), (64, 4), (128, 4), (128, 8), (256, 4), (256, 8)\}$
	Number of layers	$\{2, 3, 4\}$
	Feed-forward dimension	$\{128, 256, 512\}$
	Dropout	$\{0.1, 0.2\}$
	Pooling	$\{\text{CLS}, \text{mean}, \text{last}\}$
	Learning rate	$\{10^{-5},\ 5\times10^{-5},\ 10^{-4},\ 3\times10^{-4},\ 3\times10^{-3}\}$
	Batch size	$\{16, 32, 64\}$
	Context length	$\{64, 256, 512\}$
GRU–ODE	Embedding dimension	$\{32, 64, 128, 256\}$
	Hidden dimension	$\{32, 64, 128\}$
	ODE hidden multiplier	$\{1, 2, 4\}$
	Dropout	$\{0.1, 0.2\}$
	Reconstruction weight	$\{0, 0.05, 0.1, 0.2\}$
	Learning rate	$\{5\times10^{-6},\ 10^{-5},\ 5\times10^{-5},\ 10^{-4},\ 3\times10^{-4},\ 10^{-3}\}$
	Batch size	$\{64, 128, 256\}$
	Context length	$\{64, 256, 512\}$
DeepHit	α	$\{0.2, 0.5\}$, with ranking weight $(1-\alpha)/\alpha$
	Learning rate	$\{10^{-4}, 10^{-3}\}$
Dynamic-DeepHit	Auxiliary weight β	$\{0.01, 0.1, 0.3, 1.0\}$
	Learning rate	{selected DeepHit learning rate, 10^{-4}}
MOTOR-Probe	Batch size	$\{32, 64, 256, 512\}$
	Learning rate	$\{5\times10^{-5},\ 10^{-4},\ 10^{-3},\ 5\times10^{-3}\}$
	Gradient clipping	$\{1.0, \infty\}$

B.3 Matched input-token budget

For from-scratch encoders, we compare 3L with context length L against 1L with context length $3L$. Both settings therefore process the same nominal number of input tokens per patient and have identical parameter counts. We separately report per-example FLOPs, peak memory, and total training FLOPs. Shorter contexts reduce computation per example, while total training cost can additionally depend on the number of optimization steps required for convergence.

B.4 Landmark-count sweep

We evaluate $K \in \{1, 3, 5, 10\}$ landmarks using a fixed optimizer budget independent of K. This isolates the effect of increasing the number of prediction times from the effect of taking additional gradient steps.

C Evaluation Metrics

C.1 Standard survival performance

We report Harrell's concordance for overall discrimination and Uno's IPCW concordance at forecast horizons $h \in \{1, 3, 5, 10\}$ years. Probabilistic prediction error is measured with the IPCW Brier score and summarized with a trapezoidal integrated Brier score (IBS) over the same horizons.

For every comparison, censoring weights are estimated once on the aligned evaluation population and shared across the compared models. This prevents changes in the censoring reference from contributing to the measured 1L–3L difference.

C.2 Temporal consistency

For consecutive landmarks of the same patient separated by at least 30 days, the earlier survival prediction is transported to the later prediction time. If the landmarks are separated by Δ,

$$\widetilde{S}_i(u) = \frac{S_i(u+\Delta)}{S_i(\Delta)}. \tag{4}$$

We compare the transported prediction with the survival curve generated directly at the later landmark using total-variation distance between their induced interval probability masses:

$$m_{\mathrm{TV}} = \frac{1}{2}\sum_k |q_k - p_k|. \tag{5}$$

Lower values indicate greater temporal consistency.

A pair is included only when transport is numerically stable for both 1L and 3L, and the same patient–landmark pairs are used for both settings. This prevents a poorly behaved model from being evaluated on an easier subset of its own predictions.

C.3 Hazard dispersion

Temporal agreement can improve trivially if predicted hazards become nearly constant over forecast time. We therefore measure hazard dispersion alongside consistency. For each piecewise-exponential prediction, dispersion is the standard deviation of its log interval hazards, averaged across evaluated rows. Lower dispersion denotes a flatter hazard; unlike consistency, it has no intrinsically favorable direction. We therefore use it as a diagnostic rather than a performance metric.

C.4 Patient-level trajectory direction

For forecast horizon h, predicted risk is

$$r_{p,a}(h) = 1 - S_{p,a}(h).$$

For patients with at least three landmarks at two distinct times, we regress predicted risk on landmark calendar time separately within each patient. We report the median patient-level slope for observed cases and censored controls. A positive case slope indicates that predicted risk increases as the observed event approaches.

We additionally summarize case–control separation using

$$\mathrm{slopeAUC}(h) = \Pr\left(\widehat{b}_{\mathrm{case}} > \widehat{b}_{\mathrm{control}}\right) + \frac{1}{2}\Pr\left(\widehat{b}_{\mathrm{case}} = \widehat{b}_{\mathrm{control}}\right). \tag{6}$$

A slope AUC of 0.5 indicates no separation between case and control trajectory slopes.

C.5 Retraining reproducibility

We quantify dependence on training-fold composition using fold variance: variation in a patient's prediction across the five trained fold-models. Lower values indicate less retraining variability. Let $S_{i,a,f,h}$ denote the prediction for patient i, landmark a, fold-model f, and horizon h. With equal patient weighting,

$$V_{\mathrm{fold}} = \mathbb{E}_i\mathbb{E}_a\left[\mathrm{Var}_f(S_{i,a,f,h})\right] \tag{7}$$

C.6 Uncertainty and reporting

Five models per configuration are trained using overlapping development folds and evaluated on a fixed held-out test population. We report means and standard deviations across fold-models, paired

fold differences, and counts of folds or configurations sharing the same effect direction. Because development folds overlap substantially, these quantities should not be interpreted as independent estimates of test-set sampling uncertainty.

Multi-landmark evaluation also produces correlated rows within patients. Compared models are therefore aligned on identical patient–landmark rows. Patient-level filters are shared across configurations, and patient-cluster resampling is used where explicitly stated.

D EHR-Specific Robustness Analyses

D.1 Record truncation

We test whether the 3L advantage depends on documentation near the observed endpoint by progressively removing recent history at evaluation without retraining the models. Starting from the three-month gap already imposed during landmark construction, we evaluate total minimum separations of 3, 9, 15, and 21 months between the last code available to the model and the event or censoring time.

For each truncation level, the effective landmark, available history, patient age, and residual-time label are recomputed. We restrict all truncation levels to the population that remains eligible under the largest truncation, producing paired comparisons across levels. The censoring reference is refitted to the scored evaluation rows at each level. Because truncation changes the residual-time distribution, our estimand is the 3L–1L contrast within each level rather than absolute performance across truncation levels.

D.2 Coupling to EHR recording patterns

EHR recording intensity can carry both clinical signal and information about how care is delivered. We quantify the association of predicted risk with two properties of the observed input:

$$\text{density} = \frac{n_{\text{codes}}}{\max(\text{span in days}, 1)} \times 30.4375, \qquad \text{redundancy} = 1 - \frac{n_{\text{unique}}}{n_{\text{codes}}}. \tag{8}$$

Within patients known to remain event-free through horizon h, we compute the age-adjusted partial Spearman correlation between each recording feature and predicted log cumulative hazard. Age adjustment is performed by correlating residuals after projecting the ranked feature and ranked predicted risk on ranked age.

For 1L–3L comparisons, we report the change in absolute coupling,

$$\Delta_{\text{decouple}} = |\rho_{1\text{L}}| - |\rho_{3\text{L}}|. \tag{9}$$

Positive values indicate weaker association with the recording feature under 3L. This statistic measures statistical coupling, not causal recording bias: documentation patterns may themselves contain prognostic information.

E Additional Prediction and Efficiency Results

E.1 Absolute performance under multi-landmark evaluation

Table 5 reports the absolute 1L and 3L performance corresponding to the gain table in the main text.

E.2 Landmark-count sweep

The landmark-count sweep evaluates whether the improvement saturates as more prediction times are added under a fixed optimizer budget.

E.3 Outcome support and horizon-specific performance

Table 7 reports the number of event and at-risk landmark observations available at each forecast horizon. Figures 7 and 8 provide the task- and model-specific versions of the horizon analysis summarized in the main text.

Table 5: **Absolute predictive performance under multi-landmark evaluation.** Harrell's C-index and integrated Brier score (IBS), in percent, are reported for one-landmark (1L) and three-landmark (3L) training across all tasks and model families. Evaluation uses the ten-landmark test pool. Values are mean ± standard deviation over five training folds. Higher C-index and lower IBS indicate better performance; bold denotes the better training configuration within each task–model pair and metric. We additionally report the exploratory MOTOR-Probe 3L configuration with within-patient ranking used in the Discussion section (MOTOR-Probe + r), which achieved the strongest overall predictive performance among the evaluated MOTOR-Probe variants.

		EHRSHOT			MIMIC-IV				
Model	Landmarks	AMI	HPT	PC	BC	ADRD	DB	HPT	LC
(a) Harrell's C (↑)									
GRU	1L	59.14 ±1.94	56.15 ±1.71	60.71 ±1.59	50.95 ±3.97	61.58 ±5.33	64.64 ±0.45	70.74 ±0.21	60.31 ±14.95
GRU	3L	58.03 ±2.97	59.61 ±2.77	65.00 ±3.99	53.53 ±3.69	67.68 ±1.27	68.14 ±0.46	72.29 ±0.38	70.26 ±2.30
Transformer	1L	62.22 ±3.27	57.25 ±8.38	54.56 ±7.83	51.46 ±1.19	65.16 ±2.44	62.64 ±0.89	66.47 ±2.41	65.64 ±2.06
Transformer	3L	62.40 ±1.30	57.93 ±4.89	60.46 ±7.31	55.33 ±4.68	67.81 ±0.88	65.86 ±1.50	68.48 ±2.54	70.64 ±1.64
GRU-ODE	1L	62.13[§] ±3.75	45.79 ±4.26	58.04 ±9.87	48.66 ±0.91	59.05 ±1.47	52.32 ±2.07	64.33 ±0.33	70.32 ±2.04
GRU-ODE	3L	67.32[§] ±1.38	57.75 ±0.76	64.41 ±2.63	56.82 ±1.03	69.55 ±0.68	65.69 ±2.01	70.43 ±3.37	72.63 ±0.79
DeepHit	1L	61.23 ±0.96	55.50 ±1.50	57.48 ±2.30	52.10 ±3.18	64.82 ±2.78	66.18 ±0.91	70.23 ±0.87	66.22 ±0.34
DeepHit	3L	61.70 ±0.97	61.31 ±2.23	**68.29** ±1.40	55.15 ±2.96	70.47 ±0.66	68.33 ±0.27	71.14 ±0.70	69.21 ±1.20
Dynamic-DeepHit	1L	60.39 ±3.18	55.60 ±1.60	57.93 ±3.24	51.95 ±2.02	64.16 ±3.07	66.61 ±0.51	70.47 ±0.56	64.92 ±2.00
Dynamic-DeepHit	3L	61.15 ±2.30	59.58 ±1.77	68.10 ±1.26	56.48 ±4.00	70.10 ±1.42	68.06 ±0.61	71.10 ±1.26	69.62 ±1.44
MOTOR-Probe	1L	57.50 ±1.29	61.00 ±1.07	62.39 ±0.71	59.75 ±0.61	71.72 ±0.36	73.88 ±0.13	**72.52** ±0.16	74.56 ±0.39
MOTOR-Probe	3L	63.30 ±1.14	66.10 ±0.59	62.90 ±0.30	60.58 ±0.67	72.36 ±0.29	**74.06** ±0.20	72.45 ±0.16	75.12 ±0.17
MOTOR-Probe + r	3L	**73.96** ±0.84	**68.76** ±0.95	65.69 ±0.87	**61.46** ±0.73	**73.06** ±0.17	73.54 ±0.16	72.16 ±0.14	**76.26** ±0.26
(b) Integrated Brier score (IBS, ↓)									
GRU	1L	18.60 ±2.10	37.45 ±6.15	13.67 ±2.91	54.69 ±40.19	18.55 ±1.59	13.37 ±0.57	18.29 ±0.50	29.73 ±30.39
GRU	3L	18.32 ±2.87	24.15 ±2.45	8.85 ±2.53	40.55 ±41.25	15.32 ±0.38	11.54 ±0.20	16.43 ±0.39	11.96 ±0.39
Transformer	1L	23.23 ±6.09	68.10 ±15.22	9.86 ±4.43	14.00 ±2.55	25.00 ±8.86	13.80 ±2.11	23.67 ±2.94	13.64 ±1.36
Transformer	3L	16.89 ±4.26	55.73 ±22.47	9.21 ±3.48	11.30 ±1.09	17.51 ±0.93	12.09 ±0.51	20.92 ±2.77	12.93 ±0.71
GRU-ODE	1L	15.11[§] ±4.81	32.00 ±4.42	11.49 ±6.53	19.93 ±8.03	19.33 ±0.54	14.42 ±0.79	22.31 ±0.64	13.69 ±0.98
GRU-ODE	3L	9.21[§] ±0.51	26.08 ±1.70	10.71 ±1.49	10.15 ±0.30	15.47 ±0.40	11.98 ±0.59	17.45 ±1.29	11.67 ±0.26
DeepHit	1L	27.83 ±3.87	37.16 ±7.55	16.21 ±1.96	23.71 ±19.34	17.12 ±0.45	12.28 ±0.70	17.95 ±0.42	24.72 ±6.09
DeepHit	3L	21.26 ±11.07	20.35 ±1.78	3.60 ±0.65	10.39 ±0.68	15.30 ±0.18	11.16 ±0.13	16.61 ±0.29	11.74 ±0.14
Dynamic-DeepHit	1L	27.70 ±3.25	44.34 ±3.98	16.32 ±2.26	23.44 ±19.45	17.05 ±0.61	11.97 ±0.37	17.77 ±0.41	21.81 ±6.69
Dynamic-DeepHit	3L	17.30 ±5.21	22.51 ±2.48	3.98 ±0.88	10.68 ±0.54	15.16 ±0.16	11.25 ±0.26	16.62 ±0.27	11.80 ±0.18
MOTOR-Probe	1L	13.65 ±0.94	26.79 ±1.27	14.21 ±1.49	10.26 ±0.09	14.07 ±0.08	10.62 ±0.06	15.94 ±0.05	11.43 ±0.07
MOTOR-Probe	3L	11.34 ±0.75	22.28 ±0.66	8.24 ±0.80	9.97 ±0.04	13.92 ±0.07	10.43 ±0.05	15.65 ±0.05	11.23 ±0.04
MOTOR-Probe + r	3L	**7.88** ±0.06	**15.59** ±0.21	**2.76** ±0.06	**9.86** ±0.05	**13.49** ±0.07	**10.32** ±0.03	**15.15** ±0.06	**10.96** ±0.06

Tasks: AMI = Acute myocardial infarction, HPT = Hypertension, PC = Pancreatic cancer, BC = Breast cancer, ADRD = Alzheimer's disease and related dementia, DB = Diabetes, LC = Lung cancer. "+ r" = 3L setting and a within-patient case-ranking loss. Bold: best in each task column.

Table 6: **Effect of the number of training landmarks under latest-landmark evaluation.** Harrell's C-index and integrated Brier score (IBS), in percent, are reported as the number of training landmarks K increases. Evaluation uses the latest eligible landmark for each test patient. All configurations share the same patients and fixed optimizer budget, so L changes only the diversity of supervised prediction times. Within each metric block, bold denotes the best-performing L for each task together with configurations not distinguishable from it by the paired fold-level comparison.

		EHRSHOT			MIMIC-IV				
Metric	L	AMI	HPT	PC	BC	ADRD	DB	HPT	LC
C-idx	1	54.43 ±2.49	65.66 ±2.65	54.82 ±2.09	59.61 ±1.04	71.82 ±0.46	75.90 ±0.24	72.22 ±0.21	74.87 ±0.35
C-idx	3	61.99 ±1.55	71.73 ±2.09	56.61 ±2.25	**60.68** ±1.29	72.83 ±0.31	75.91 ±0.22	72.31 ±0.18	**75.45** ±0.49
C-idx	5	63.51 ±1.68	71.32 ±2.24	58.31 ±3.17	**60.90** ±1.15	72.94 ±0.36	**76.11** ±0.18	**72.36** ±0.15	**75.74** ±0.30
C-idx	10	**66.03** ±1.62	**72.10** ±1.86	**60.38** ±2.99	60.52 ±1.21	**73.01** ±0.34	76.02 ±0.17	72.27 ±0.16	**75.57** ±0.46
IBS	1	11.52 ±0.51	20.70 ±0.86	5.23 ±0.29	13.26 ±0.11	15.65 ±0.20	11.64 ±0.03	16.58 ±0.06	13.99 ±0.14
IBS	3	10.07 ±0.36	**18.87** ±0.30	**4.34** ±0.19	**13.00** ±0.09	15.75 ±0.15	11.72 ±0.06	16.37 ±0.07	14.08 ±0.21
IBS	5	10.13 ±0.37	**18.98** ±0.37	**4.34** ±0.17	**12.94** ±0.06	**15.54** ±0.12	**11.58** ±0.05	**16.36** ±0.07	**13.92** ±0.18
IBS	10	**9.94** ±0.33	**18.91** ±0.38	**4.33** ±0.15	**12.95** ±0.04	15.59 ±0.09	11.66 ±0.06	**16.35** ±0.06	**13.94** ±0.20

Tasks: AMI = Acute myocardial infarction, HPT = Hypertension, PC = Pancreatic cancer, BC = Breast cancer, ADRD = Alzheimer's disease and related dementia, DB = Diabetes, LC = Lung cancer.

E.4 MATCHED-BUDGET CONTROLS

The matched-token and matched-update analyses test whether the main gain can be explained by additional input or optimizer steps.

Table 7: **Outcome support across forecast horizons under multi-landmark evaluation.** Number of test patient–landmark pairs with an observed event within each forecast horizon (*Events*) or remaining observed and event-free through that horizon (*At risk*). Landmarks censored before the corresponding horizon contribute to neither count. Counts are reported for the ten-landmark test pool used in the horizon-specific performance analyses.

			Horizon			
Dataset	Task		1 y	3 y	5 y	10 y
EHRSHOT	AMI	Events	292	534	664	845
		At risk	7,845	5,470	3,819	1,349
	HPT	Events	372	704	904	1,146
		At risk	4,342	2,917	1,939	636
	PC	Events	80	153	189	241
		At risk	8,086	5,717	4,004	1,447
MIMIC-IV	BC	Events	121	311	436	609
		At risk	5,199	3,649	2,350	508
	ADRD	Events	459	1,095	1,549	2,036
		At risk	7,316	4,831	3,115	741
	DB	Events	604	1,278	1,691	2,128
		At risk	13,574	9,769	6,587	1,570
	HPT	Events	1,323	2,575	3,284	4,122
		At risk	10,882	7,328	4,884	981
	LC	Events	276	646	816	1,111
		At risk	7,030	4,493	2,827	515

Tasks: AMI = Acute myocardial infarction, HPT = Hypertension, PC = Pancreatic cancer, BC = Breast cancer, ADRD = Dementia, DB = Diabetes, LC = Lung cancer.

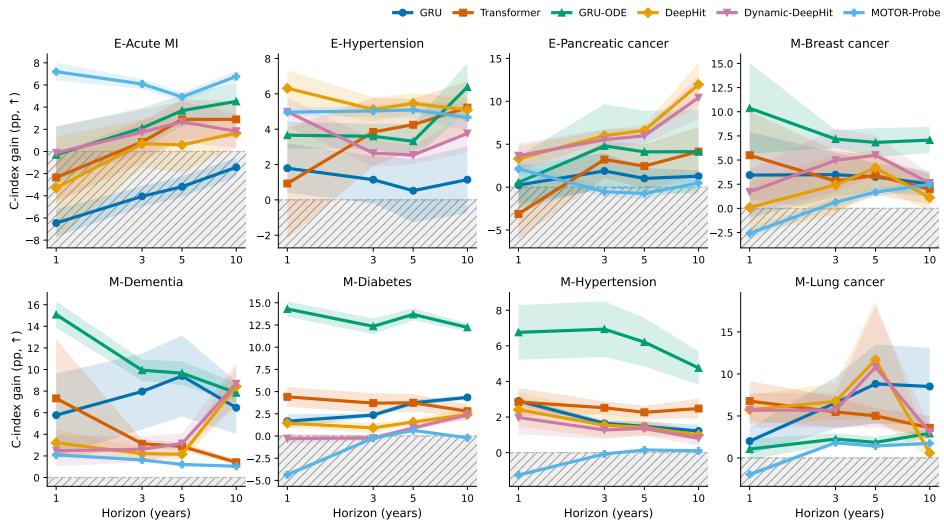

Figure 7: **Horizon-specific discrimination gains by task and model family.** Difference in Uno's concordance between three-landmark and one-landmark training across 1-, 3-, 5-, and 10-year forecast horizons, shown separately for each task and model family. Values are averaged over five training folds. Positive values indicate better discrimination with three-landmark training.

E.5 Training-cohort size

We additionally reduce the number of training patients available to the MOTOR-Probe while retaining three landmarks per patient. Figure 9 compares each 3L configuration against the 1L model trained on the complete cohort.

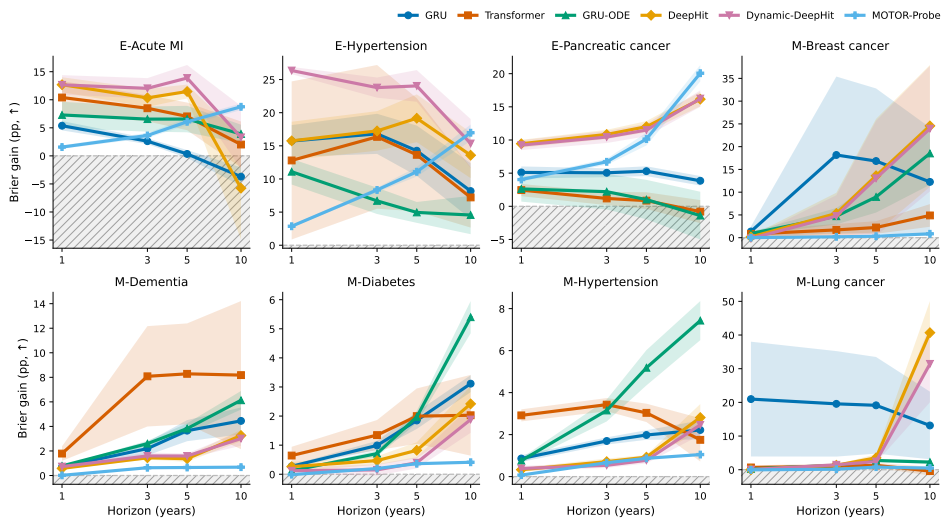

Figure 8: **Horizon-specific Brier-score gains by task and model family.** Difference in horizon-specific Brier score between three-landmark and one-landmark training across 1-, 3-, 5-, and 10-year forecast horizons, shown separately for each task and model family. Values are averaged over five training folds and oriented so that positive values indicate lower prediction error with three-landmark training.

Table 8: **Three-landmark training improves prediction under a matched input-token budget.** We compare one landmark with context length $3M$ against three landmarks with context length M, matching the nominal number of input tokens per patient and model parameter count. Performance is evaluated on the ten-landmark test pool. Gains are paired within fold and reported in percentage points as $C_{3L} - C_{1L}$ and $\mathrm{IBS}_{1L} - \mathrm{IBS}_{3L}$, so positive values favor three-landmark training. Computational ratios are reported as 1L÷3L; values above one indicate lower cost for 3L. Forward FLOPs measure cost per training example, total training FLOPs additionally account for the number of examples processed before early stopping, and peak memory is measured at batch size 16 on the same GPU. Values are mean ± standard deviation over five training folds.

				Gain of 3 landmarks (pp)		Cost ratio 1 / 3 landmarks		
Dataset	Task	Model	Context (landmarks × codes)	Harrell's C	IBS	FLOPs / sample	Training FLOPs	Memory
EHRSHOT	AMI	GRU	1×1536 vs. 3×512	−2.55 ±4.63	2.72 ±2.37	3.00×	1.22× ±0.43	2.79×
		Transformer	1×1536 vs. 3×512	−2.70 ±3.83	1.68 ±5.82	4.46×	1.33× ±0.13	8.17×
		GRU-ODE	1×768 vs. 3×256	6.10 ±3.29	10.72 ±8.60	3.00×	1.41× ±0.61	2.86×
	HPT	GRU	1×1536 vs. 3×512	3.09 ±2.98	11.67 ±6.79	3.00×	1.28× ±0.55	2.39×
		Transformer	1×768 vs. 3×256	0.87 ±5.74	8.15 ±8.86	3.97×	1.21× ±0.33	5.18×
		GRU-ODE	1×192 vs. 3×64	9.22 ±5.80	5.62 ±5.78	3.00×	0.93× ±0.60	2.13×
	PC	GRU	1×768 vs. 3×256	5.84 ±4.79	9.35 ±2.25	3.00×	0.80× ±0.46	2.05×
		Transformer	1×768 vs. 3×256	5.92 ±5.01	2.28 ±1.45	4.16×	2.30× ±1.18	6.63×
		GRU-ODE	1×768 vs. 3×256	6.75 ±3.13	4.64 ±2.99	3.00×	1.25× ±0.61	2.85×
MIMIC-IV	BC	GRU	1×512 vs. 3×170	2.52 ±6.70	14.11 ±29.80	3.01×	1.02× ±0.54	1.70×
		Transformer	1×512 vs. 3×170	1.00 ±4.81	4.81 ±1.96	4.07×	1.56× ±0.47	5.00×
		GRU-ODE	1×512 vs. 3×170	7.97 ±1.73	9.77 ±8.19	3.01×	0.71× ±0.84	2.84×
	ADRD	GRU	1×512 vs. 3×170	4.94 ±3.22	14.30 ±25.39	3.01×	0.86× ±0.33	1.71×
		Transformer	1×512 vs. 3×170	1.08 ±2.19	7.21 ±9.35	3.61×	1.64× ±0.57	6.25×
		GRU-ODE	1×512 vs. 3×170	12.30 ±2.48	4.98 ±0.83	3.01×	0.34× ±0.01	2.50×
	DB	GRU	1×512 vs. 3×170	2.87 ±0.29	1.79 ±0.57	3.01×	1.09× ±0.15	1.57×
		Transformer	1×512 vs. 3×170	−0.02 ±1.70	0.94 ±1.05	4.06×	1.50× ±0.55	6.70×
	HPT	GRU	1×512 vs. 3×170	0.86 ±0.20	1.98 ±0.35	3.01×	1.69× ±0.38	1.81×
		Transformer	1×512 vs. 3×170	1.40 ±0.47	3.02 ±0.58	3.61×	2.17× ±0.38	6.19×
		GRU-ODE	1×768 vs. 3×256	6.18 ±3.10	5.28 ±2.07	3.00×	0.69× ±0.28	2.26×
	LC	GRU	1×512 vs. 3×170	9.36 ±15.57	10.61 ±21.84	3.01×	1.34× ±0.76	1.60×
		Transformer	1×512 vs. 3×170	2.45 ±2.13	1.16 ±0.95	3.71×	1.27× ±0.24	5.53×
		GRU-ODE	1×512 vs. 3×170	0.68 ±1.34	1.68 ±0.56	3.01×	1.32× ±0.29	2.59×

Tasks: AMI = Acute myocardial infarction, HPT = Hypertension, PC = Pancreatic cancer, BC = Breast cancer, ADRD = Alzheimer's disease and related dementia, DB = Diabetes, LC = Lung cancer.

F Additional Robustness Results

F.1 Coupling to recording patterns

Table 10 provides task- and model-specific changes in coupling to code density and redundancy.

Table 9: **Three-landmark supervision improves MOTOR-Probe performance at a fixed optimizer budget.** One-landmark (1L) and three-landmark (3L) MOTOR-Probe models are trained on the same patients for the same number of optimizer updates and evaluated on the ten-landmark test pool. We report mean ± standard deviation over five training folds for Harrell's C-index and integrated Brier score (IBS), in percentages. Higher C-index and lower IBS indicate better performance; bold indicates the better training configuration for each task and metric.

			C-index (↑)		IBS (↓)	
Dataset	Task	Updates	1L	3L	1L	3L
EHRSHOT	AMI	6,000	57.50 ±1.29	**63.30** ±1.14	13.65 ±0.94	**11.34** ±0.75
	HPT	3,600	61.00 ±1.07	**66.10** ±0.59	26.79 ±1.27	**22.28** ±0.66
	PC	6,000	62.39 ±0.71	**62.90** ±0.30	14.21 ±1.49	**8.24** ±0.80
MIMIC-IV	BC	14,400	59.75 ±0.61	**60.58** ±0.67	10.26 ±0.09	**9.97** ±0.04
	ADRD	21,400	71.72 ±0.36	**72.36** ±0.29	14.07 ±0.08	**13.92** ±0.07
	DB	37,200	73.88 ±0.13	**74.06** ±0.20	10.62 ±0.06	**10.43** ±0.05
	HPT	31,800	**72.52** ±0.16	72.45 ±0.16	15.94 ±0.05	**15.65** ±0.05
	LC	19,800	74.56 ±0.39	**75.12** ±0.17	11.43 ±0.07	**11.23** ±0.04

Tasks: AMI = Acute myocardial infarction, HPT = Hypertension, PC = Pancreatic cancer, BC = Breast cancer, ADRD = Alzheimer's disease and related dementia, DB = Diabetes, LC = Lung cancer.

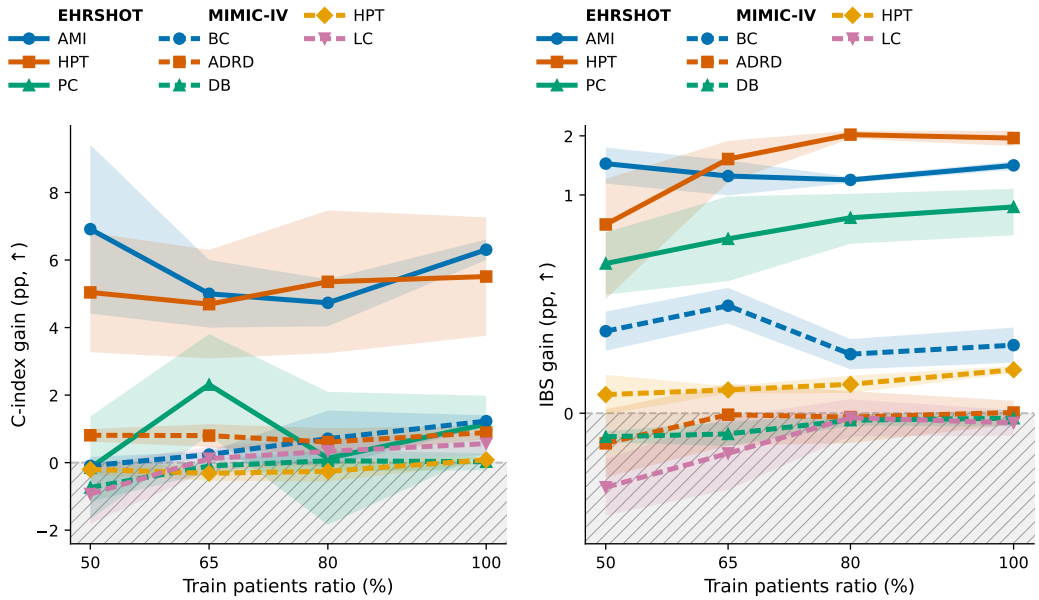

Figure 9: **Effect of training-cohort size under three-landmark supervision.** Harrell's C-index (left) and IBS (right) difference between three-landmark MOTOR-Probe models trained on increasing fractions of the training cohort and the one-landmark baseline trained on the full cohort. Positive values indicate better performance than the full-cohort 1L baseline. Results are averaged over five training folds.

F.2 Record truncation

Figures 10 and 11 decompose the truncation analysis by health system, task, and model family.

F.3 Reproducibility across retraining

The following table reports the temporal-dynamics estimate of prediction variability attributable to retraining.

Table 10: **Change in coupling between predicted risk and EHR recording patterns under three-landmark training.** For each task and model family, we report the change in the absolute age-adjusted partial Spearman correlation between predicted log cumulative hazard and either code density or code redundancy, relative to the corresponding one-landmark model. Values are averaged across five folds and forecast horizons $h \in \{1, 3, 5, 10\}$ years. Positive values indicate weaker coupling with three-landmark training. Correlations are computed among patients known to remain event-free through the corresponding horizon.

	EHRSHOT			MIMIC-IV				
Model	AMI	HPT	PC	BC	ADRD	DB	HPT	LC
Code density								
GRU	−0.141	0.073	−0.006	0.005	−0.057	0.039	−0.026	0.046
Transformer	−0.148	0.055	−0.164	−0.025	0.006	0.025	−0.060	0.037
GRU-ODE	–	0.069	0.015	–	0.225	–	0.257	–
DeepHit	−0.107	0.073	0.113	0.046	0.010	0.020	−0.001	0.035
Dynamic-DeepHit	−0.043	0.032	0.120	0.043	0.023	0.016	−0.005	0.015
MOTOR-Probe	0.053	0.032	0.002	0.022	0.026	0.033	0.027	0.023
Code redundancy								
GRU	0.000	0.074	0.079	−0.044	−0.041	0.102	−0.038	0.018
Transformer	−0.017	0.008	−0.019	−0.009	0.016	−0.024	−0.060	0.026
GRU-ODE	–	0.081	0.108	–	0.361	–	0.006	–
DeepHit	0.165	0.165	0.057	0.105	−0.008	−0.030	−0.070	−0.036
Dynamic-DeepHit	0.187	0.115	0.083	0.105	−0.026	−0.025	−0.089	−0.070
MOTOR-Probe	0.037	0.091	0.048	0.049	0.038	0.005	−0.057	0.014

Tasks: AMI = Acute myocardial infarction, HPT = Hypertension, PC = Pancreatic cancer, BC = Breast cancer, ADRD = Alzheimer's disease and related dementia, DB = Diabetes, LC = Lung cancer.

G Additional Temporal-Dynamics Results

G.1 Temporal consistency and hazard dispersion

Table 12 reports temporal-consistency error and hazard dispersion for each task–model combination. Hazard dispersion is shown as a diagnostic and is not assigned a preferred direction.

G.2 Patient-level risk trajectories

Table 13 reports median case and control risk slopes and slope AUC at the 5-year horizon.

H Exploratory Within-Patient Ordering

The trajectory analysis in the main paper distinguishes supervision at multiple prediction times from explicit supervision of relationships between those times. As a proof of principle, we therefore evaluate a training configuration that introduces within-patient ordering.

H.1 Training configuration

Landmarks belonging to the same patient are grouped within a batch. For observed cases, landmarks are ordered in calendar time and consecutive valid pairs are compared. The trajectory objective encourages later one-year cumulative hazard to exceed earlier cumulative hazard by margin m:

$$\mathcal{L}_{\text{traj}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{1}{|\mathcal{Q}_p|} \sum_{(a,a') \in \mathcal{Q}_p} \max\left\{0, H_{p,a}(1) - H_{p,a'}(1) + m\right\}, \tag{10}$$

with $m = 0.05$. Controls do not contribute to this ordering loss.

The total training objective combines the survival and trajectory terms with unit trajectory-loss weight. This exploratory configuration also uses patient-grouped batches, explicit age conditioning through FiLM, and a different checkpoint-selection criterion from the plain 3L configuration.

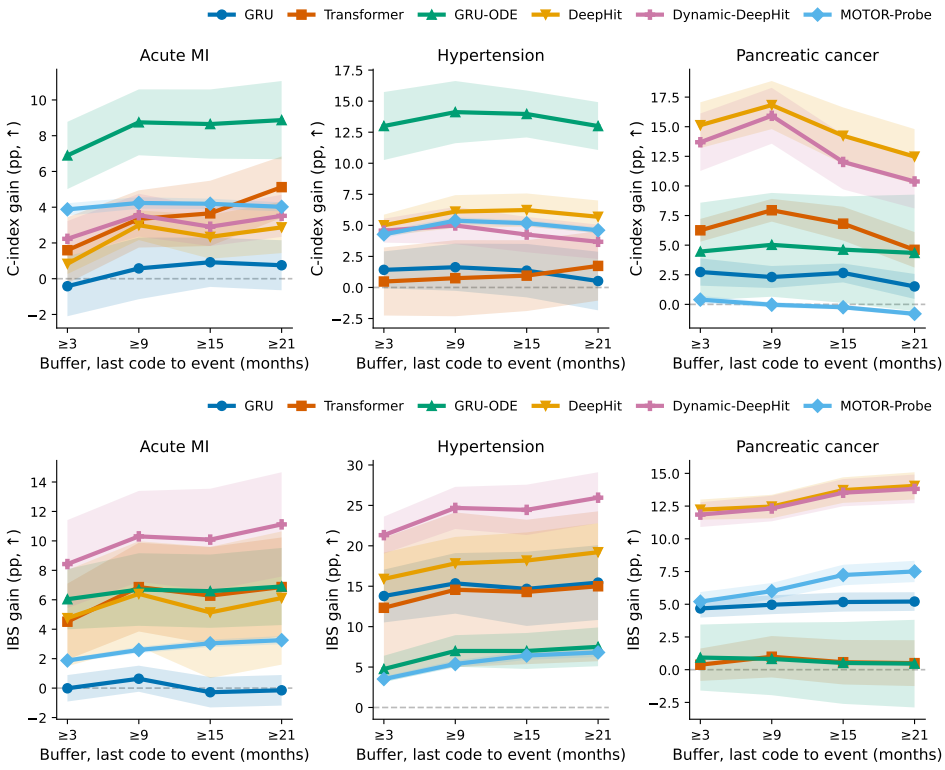

Figure 10: **Prediction gains persist under increasing record truncation on EHRSHOT.** Difference between three-landmark and one-landmark training in Harrell's C-index (top) and IBS (bottom), shown separately by task as the minimum interval between the last observed code and the event or censoring time is increased at evaluation. All models are trained with the same three-month minimum gap and are not retrained across truncation levels. Positive values indicate better performance with 3L.

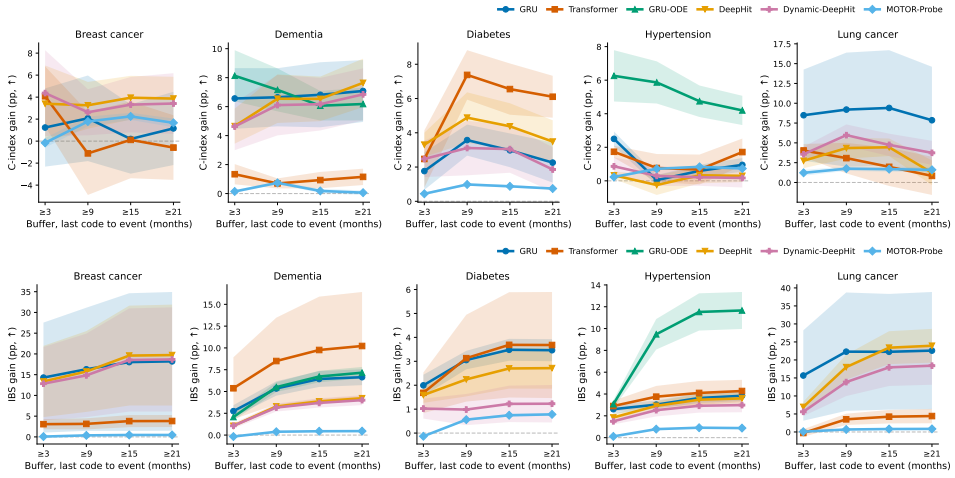

Figure 11: **Prediction gains persist under increasing record truncation on MIMIC-IV.** Difference between three-landmark and one-landmark training in Harrell's C-index (top) and IBS (bottom), shown separately by task as the minimum interval between the last observed code and the event or censoring time is increased at evaluation. All models are trained with the same three-month minimum gap and are not retrained across truncation levels. Positive values indicate better performance with 3L.

We therefore interpret the comparison as a recipe-level proof of principle rather than an isolated estimate of the ranking term.

Table 11: **Retraining variability at the 5-year horizon under one- and three-landmark training.** V_{fold} measures the variance of predicted survival across the five fold-models, averaged over landmarks within patient and then across patients; lower values indicate greater reproducibility across retraining. Bold denotes the lower-variance configuration within each task–model pair. Three-landmark training reduces V_{fold} in 26 of 31 configurations.

Dataset	Task	Model	Vfold 1L ($\times 10^{-3}$)	Vfold 3L ($\times 10^{-3}$)
EHRSHOT	Acute MI	GRU	14.6	**13.5**
EHRSHOT	Acute MI	Transformer	46.6	**8.4**
EHRSHOT	Acute MI	MOTOR-Probe	1.9	**1.08**
EHRSHOT	Hypertension	GRU	82.2	**30.3**
EHRSHOT	Hypertension	Transformer	**108**	178
EHRSHOT	Hypertension	GRU-ODE	24.7	**18.8**
EHRSHOT	Hypertension	MOTOR-Probe	2.65	**1.92**
EHRSHOT	Pancreatic cancer	GRU	38.5	**8.61**
EHRSHOT	Pancreatic cancer	Transformer	34.1	**9.19**
EHRSHOT	Pancreatic cancer	GRU-ODE	15.8	**10.9**
EHRSHOT	Pancreatic cancer	MOTOR-Probe	1.28	**0.585**
MIMIC-IV	Breast cancer	GRU	**193**	221
MIMIC-IV	Breast cancer	Transformer	29.7	**17.9**
MIMIC-IV	Breast cancer	GRU-ODE	26.2	**2.08**
MIMIC-IV	Breast cancer	MOTOR-Probe	0.759	**0.523**
MIMIC-IV	Dementia	GRU	11.4	**10.8**
MIMIC-IV	Dementia	Transformer	84.3	**25.6**
MIMIC-IV	Dementia	GRU-ODE	**2.01**	19
MIMIC-IV	Dementia	MOTOR-Probe	1.22	**0.81**
MIMIC-IV	Diabetes	GRU	9.36	**6.61**
MIMIC-IV	Diabetes	Transformer	22.6	**11.5**
MIMIC-IV	Diabetes	GRU-ODE	**0.901**	9.27
MIMIC-IV	Diabetes	MOTOR-Probe	0.568	**0.387**
MIMIC-IV	Hypertension	GRU	16.4	**14.3**
MIMIC-IV	Hypertension	Transformer	62.2	**44.2**
MIMIC-IV	Hypertension	GRU-ODE	**1.51**	17.1
MIMIC-IV	Hypertension	MOTOR-Probe	0.966	**0.717**
MIMIC-IV	Lung cancer	GRU	110	**6.79**
MIMIC-IV	Lung cancer	Transformer	20.1	**11**
MIMIC-IV	Lung cancer	GRU-ODE	14.8	**9.57**
MIMIC-IV	Lung cancer	MOTOR-Probe	0.976	**0.624**

H.2 Trajectory separation

Table 12 reports slope AUC for the one-landmark, three-landmark, and ranking-augmented MOTOR-Probe configurations on the three EHRSHOT tasks.

Figure 12: **Within-patient risk direction on EHRSHOT under one-landmark, three-landmark, and ranking-augmented training.** Slope AUC at the 5-year forecast horizon for MOTOR-Probe across the three EHRSHOT tasks. Slope AUC is the probability that a case has a larger patient-level risk slope than a control; 0.5 indicates no case–control separation and higher values indicate stronger separation in trajectory direction. Values are mean ± standard deviation over five training folds. Bold indicates the highest value within each task.

Task	Baseline 1L	3L	3L + rank loss
AMI	0.609 $\pm$0.032	0.650 $\pm$0.024	**0.668** $\pm$0.009
HPT	0.546 $\pm$0.010	0.586 $\pm$0.013	**0.656** $\pm$0.012
PC	0.493 $\pm$0.029	0.542 $\pm$0.029	**0.611** $\pm$0.012

Table 12: **Temporal consistency and hazard dispersion under one- and three-landmark training.** Temporal consistency is measured by the total-variation distance (TV) between an earlier survival prediction transported to the next landmark and the prediction made at that landmark; lower values indicate greater consistency. Hazard dispersion (D_{haz}) measures variation in the predicted log interval hazards; lower values indicate smoother, flatter hazards and are not intrinsically better. Metrics are computed on identical consecutive landmark pairs for 1L and 3L and reported as mean ± standard deviation over five training folds. Bold indicates the more temporally consistent configuration within each task–model pair.

Dataset	Task	Model	TV 1L	TV 3L	Dhaz 1L	Dhaz 3L
EHRSHOT	AMI	GRU	0.216 ±0.041	**0.107** ±0.037	1.513 ±0.139	0.874 ±0.035
EHRSHOT	AMI	Transformer	0.210 ±0.089	**0.093** ±0.028	2.818 ±0.169	0.967 ±0.343
EHRSHOT	AMI	GRU-ODE[§]	0.172 ±0.058	**0.097** ±0.027	2.319 ±0.245	0.425 ±0.066
EHRSHOT	AMI	MOTOR probe	0.180 ±0.022	**0.169** ±0.009	1.372 ±0.192	0.996 ±0.038
EHRSHOT	HPT	GRU	0.288 ±0.036	**0.193** ±0.023	2.563 ±0.398	0.723 ±0.143
EHRSHOT	HPT	Transformer	0.355 ±0.130	**0.257** ±0.047	3.507 ±1.261	2.446 ±1.363
EHRSHOT	HPT	GRU-ODE	0.244 ±0.048	**0.207** ±0.027	1.237 ±0.197	0.549 ±0.029
EHRSHOT	HPT	MOTOR probe	0.241 ±0.011	**0.207** ±0.007	0.905 ±0.056	0.789 ±0.050
EHRSHOT	PC	GRU	0.186 ±0.028	**0.069** ±0.016	1.843 ±0.439	0.827 ±0.068
EHRSHOT	PC	Transformer	0.197 ±0.091	**0.059** ±0.010	2.446 ±0.347	1.625 ±0.248
EHRSHOT	PC	GRU-ODE	0.101 ±0.046	**0.065** ±0.018	2.226 ±0.709	0.862 ±0.138
EHRSHOT	PC	MOTOR probe	0.170 ±0.015	**0.143** ±0.011	1.358 ±0.112	1.139 ±0.042
MIMIC-IV	BC	GRU	0.223 ±0.343	**0.099** ±0.099	1.727 ±1.388	1.073 ±1.043
MIMIC-IV	BC	Transformer	0.207 ±0.080	**0.114** ±0.038	1.098 ±0.629	0.973 ±0.421
MIMIC-IV	BC	GRU-ODE	0.071 ±0.038	**0.059** ±0.013	0.331 ±0.114	0.269 ±0.035
MIMIC-IV	BC	MOTOR probe	0.096 ±0.005	**0.077** ±0.002	0.644 ±0.048	0.524 ±0.033
MIMIC-IV	ADRD	GRU	**0.090** ±0.027	0.110 ±0.011	0.361 ±0.131	0.320 ±0.068
MIMIC-IV	ADRD	Transformer	0.341 ±0.144	**0.242** ±0.024	1.580 ±1.176	0.975 ±0.142
MIMIC-IV	ADRD	GRU-ODE	**0.057** ±0.030	0.147 ±0.030	0.267 ±0.039	0.481 ±0.083
MIMIC-IV	ADRD	MOTOR probe	0.137 ±0.010	**0.123** ±0.002	0.627 ±0.053	0.525 ±0.033
MIMIC-IV	DB	GRU	**0.104** ±0.008	0.105 ±0.005	0.330 ±0.021	0.300 ±0.021
MIMIC-IV	DB	Transformer	0.203 ±0.091	**0.138** ±0.027	0.737 ±0.228	0.437 ±0.132
MIMIC-IV	DB	GRU-ODE	**0.078** ±0.020	0.103 ±0.014	0.268 ±0.030	0.366 ±0.023
MIMIC-IV	DB	MOTOR probe	0.107 ±0.004	**0.090** ±0.003	0.572 ±0.042	0.519 ±0.043
MIMIC-IV	HPT	GRU	0.139 ±0.010	**0.123** ±0.017	0.410 ±0.029	0.429 ±0.101
MIMIC-IV	HPT	Transformer	0.347 ±0.037	**0.260** ±0.037	1.267 ±0.118	0.957 ±0.053
MIMIC-IV	HPT	GRU-ODE	0.110 ±0.011	**0.105** ±0.013	0.272 ±0.024	0.358 ±0.048
MIMIC-IV	HPT	MOTOR probe	0.137 ±0.007	**0.121** ±0.003	0.540 ±0.035	0.472 ±0.018
MIMIC-IV	LC	GRU	0.160 ±0.148	**0.060** ±0.029	1.045 ±0.945	0.355 ±0.072
MIMIC-IV	LC	Transformer	0.170 ±0.066	**0.137** ±0.040	1.006 ±0.201	0.777 ±0.242
MIMIC-IV	LC	GRU-ODE	0.128 ±0.017	**0.115** ±0.017	0.284 ±0.019	0.461 ±0.094
MIMIC-IV	LC	MOTOR probe	0.112 ±0.009	**0.093** ±0.001	0.642 ±0.067	0.561 ±0.058

Pairs are at least 30 days apart, and the stability filter is applied to the intersection of pairs valid under both 1L and 3L; Tasks: AMI = Acute myocardial infarction, HPT = Hypertension, PC = Pancreatic cancer, BC = Breast cancer, ADRD = Alzheimer's disease and related dementia, DB = Diabetes, LC = Lung cancer.

Table 13: **Within-patient risk direction under one- and three-landmark training.** At the 5-year forecast horizon, predicted risk is regressed on landmark time separately for each patient. Case and control columns report the median patient-level slope, in 10^{-3} risk per year; positive case slopes correspond to increasing predicted risk as the observed event approaches. Slope AUC is the probability that a case has a larger slope than a control, with 0.5 indicating no case–control separation. Metrics are computed on identical patients for 1L and 3L and aggregated over five training folds. Bold indicates slope AUC above 0.5; † denotes tasks with fewer than 50 observed cases.

Dataset	Task	Model	Case 1L	Case 3L	Ctrl 1L	Ctrl 3L	AUC 1L	AUC 3L	Cases
EHRSHOT	AMI	GRU	-17.0	-15.0	-10.5	-5.0	0.479	0.463	122
EHRSHOT	AMI	Transformer	-28.8	-21.6	-24.9	-3.8	0.500	0.430	122
EHRSHOT	AMI	MOTOR-Probe	-4.7	+3.2	-4.8	-2.3	**0.561**	**0.596**	122
EHRSHOT	HPT	GRU	-16.0	-10.1	-18.4	-3.9	**0.501**	0.470	168
EHRSHOT	HPT	Transformer	-11.5	-4.1	-5.1	-4.8	0.450	0.443	168
EHRSHOT	HPT	GRU-ODE	-15.5	-7.8	-5.8	-2.4	0.434	0.482	168
EHRSHOT	HPT	MOTOR-Probe	-8.0	+1.2	-15.8	-10.2	**0.534**	**0.551**	168
EHRSHOT	PC	GRU	-34.2	-4.3	-15.4	-1.8	0.458	0.488	36†
EHRSHOT	PC	Transformer	-12.7	-3.6	-5.8	-0.5	0.467	**0.501**	36†
EHRSHOT	PC	GRU-ODE	-19.9	-1.5	-6.0	-2.4	0.490	**0.533**	36†
EHRSHOT	PC	MOTOR-Probe	-61.7	-35.3	-10.8	-7.2	0.432	0.448	36†
MIMIC-IV	BC	GRU	-2.8	-38.4	-1.1	-5.2	0.453	0.442	65
MIMIC-IV	BC	Transformer	-11.4	-3.7	-1.7	-1.6	0.488	0.488	65
MIMIC-IV	BC	GRU-ODE	-8.1	-55.9	-3.5	-8.4	0.415	0.431	65
MIMIC-IV	BC	MOTOR-Probe	-41.2	-38.0	-13.8	-9.4	0.432	0.414	65
MIMIC-IV	ADRD	GRU	-3.6	+2.4	-3.3	+0.0	**0.513**	**0.534**	218
MIMIC-IV	ADRD	Transformer	-8.7	+14.0	-5.6	+4.5	0.496	**0.528**	218
MIMIC-IV	ADRD	GRU-ODE	-11.2	+4.0	-8.2	+3.3	0.479	**0.519**	218
MIMIC-IV	ADRD	MOTOR-Probe	-23.6	-6.3	-6.0	+0.2	**0.505**	**0.519**	218
MIMIC-IV	DB	GRU	-6.0	-7.4	-8.3	-6.5	**0.527**	**0.524**	224
MIMIC-IV	DB	Transformer	-15.4	+4.7	-13.5	-1.8	**0.510**	**0.526**	224
MIMIC-IV	DB	GRU-ODE	-6.7	+8.9	-15.1	+0.6	0.485	**0.520**	224
MIMIC-IV	DB	MOTOR-Probe	-42.9	-8.2	-30.8	-18.6	**0.522**	**0.565**	224
MIMIC-IV	HPT	GRU	-41.5	-21.0	-21.4	-9.0	**0.514**	**0.512**	430
MIMIC-IV	HPT	Transformer	-55.5	-1.5	-23.2	-4.2	0.498	**0.514**	430
MIMIC-IV	HPT	GRU-ODE	-209.5	-118.3	-43.5	-15.2	0.431	0.459	430
MIMIC-IV	HPT	MOTOR-Probe	-169.4	-77.6	-40.5	-23.1	**0.500**	**0.509**	430
MIMIC-IV	LC	GRU	-202.8	-105.8	-2.9	-0.0	0.434	0.389	116
MIMIC-IV	LC	Transformer	+15.2	-1.5	+4.5	+0.8	**0.559**	0.460	116
MIMIC-IV	LC	GRU-ODE	-12.2	-165.4	-0.6	-0.1	0.478	0.381	116
MIMIC-IV	LC	MOTOR-Probe	-45.3	-28.6	-12.6	-7.6	0.478	0.476	116

Tasks: AMI = Acute myocardial infarction, HPT = Hypertension, PC = Pancreatic cancer, BC = Breast cancer, ADRD = Alzheimer's disease and related dementia, DB = Diabetes, LC = Lung cancer.