INNOPOLIS UNIVERSITY

Similar Scores, Different High-Risk Episodes: Auditing Model Selection in Longitudinal EHRs

Polina Korobeinikova (1) Dmitry Lvov (1, 2) Ilya Pershin (1, 3)

1 Innopolis University 2 ITMO University 3 Kazan Federal University

TAE (Trust-AI-Eval), NeurIPS 2026 d.lvov@innopolis.ru

Research Center of the Artificial Intelligence Institute, Innopolis University, Innopolis, Russia

Model selection at a fixed review capacity

Clinical risk scores become finite priority lists.

Cohort-level performance

AUROC, average precision and probability losses summarize predictions across the held-out cohort.

High-risk prioritization

The highest-risk 10% form a discrete set. Changes in local ranking can change which episodes enter.

Same capacity

k = ceil(0.10 N) episodes

Same selection?

Compare membership as well as scores

Do metric rankings, prioritized sets and behavioral tests support the same representation choice?

EHRSHOT cohorts and common protocol

30-day readmission and ICU transfer, with patient-disjoint train, tuning and held-out splits.

Held-out cohorts

Task	Episodes	Patients	Positives
Readmission	2,189	1,190	260
ICU transfer	2,037	1,154	85

Readmission: numeric RETAIN-lite

ICU transfer: numeric two-layer GRU

Train

Models, persistent-code list, vocabulary and normalization

Tuning

Checkpoint selection and seed-wise Platt calibration

Held-out

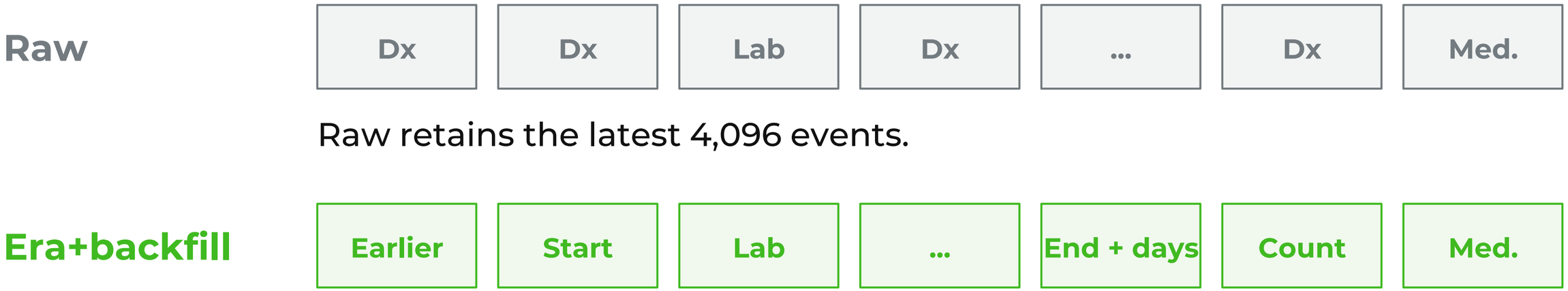
Average five calibrated predictions, then compare pipelines

Prediction at 23:59: discharge day for readmission, admission day for ICU transfer. Seeds 42-46.

Raw history and persistence-aware state

Both pipelines use a 4,096-event budget. The train-only rule selects 239 persistent diagnosis codes.

Schematic: Dx = selected diagnosis, Lab = measurement, Med. = medication



Consecutive visit-day mentions of a selected diagnosis form an era when each gap is at most 90 days. Transform the complete history before applying the 4,096-event cap. State tokens encode start, duration and count.

The Raw comparison changes explicit state encoding and the history admitted within the same event budget.

Primary five-seed ensemble performance

Era denotes Era+backfill at 4,096 events. Higher AUROC, AP and precision are better. Lower losses are better.

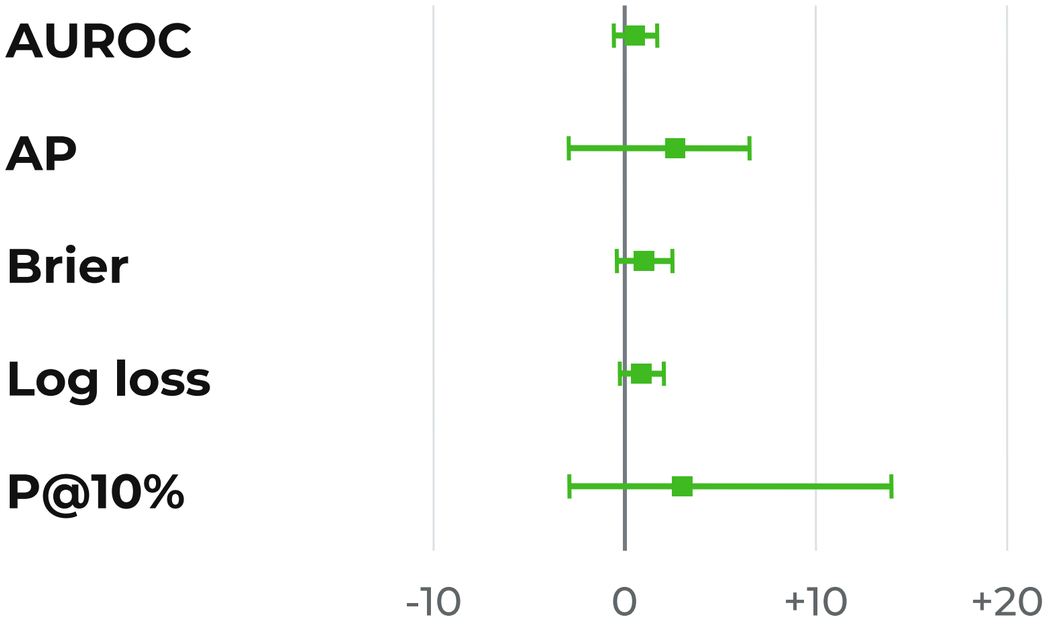
	Readmission		ICU transfer	
Metric	**Raw**	**Era**	**Raw**	**Era**
AUROC	0.7942	0.7985	0.7948	0.8013
AP	0.4097	0.4206	0.2194	0.2241
Brier score	0.08690	0.08602	0.03637	0.03618
Log loss	0.29956	0.29693	0.14926	0.14745
P@10%	0.4521	0.4658	0.1765	0.1814

Era leads all ten primary point estimates. The paired uncertainty intervals appear on the next slide.

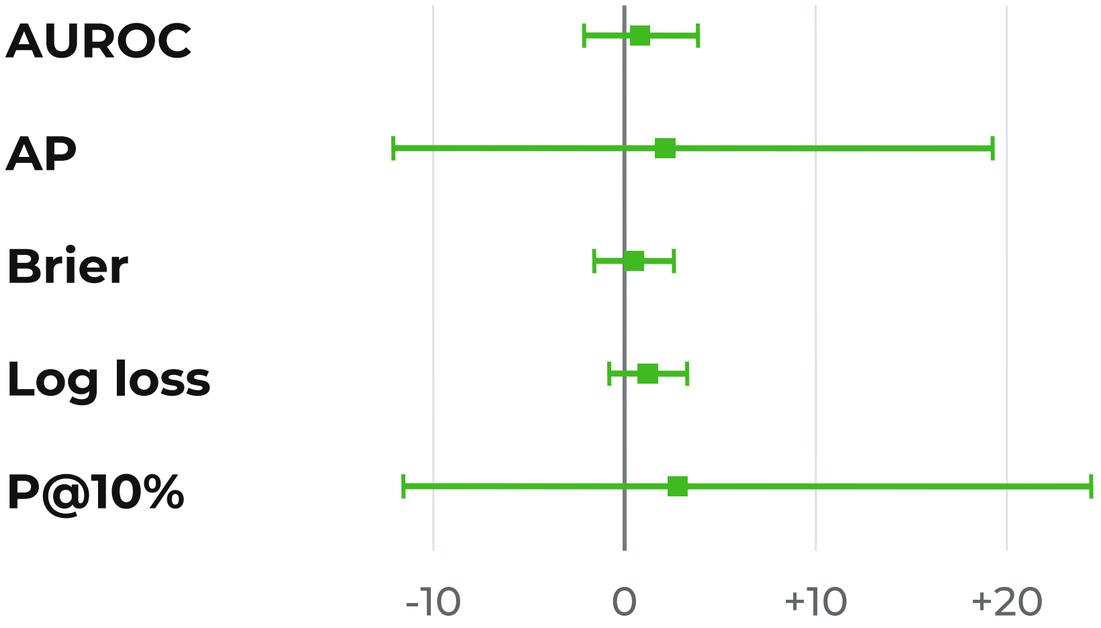
Paired uncertainty in the primary comparison

Relative benefit as a percentage of the Raw metric. Positive favors Era. Loss directions are reversed.

Readmission



ICU transfer



All ten 95% intervals include zero. 10,000 paired patient-cluster draws, conditional on the fitted ensembles.

Different episodes fill the same capacity

Each list selects ceil(0.10 N) episodes. Replacement is the fraction of each list absent from the other.

Readmission



185 shared; 34 of 219 specific to each list

15.5%

Captured positives
Raw 99 Era 102

ICU transfer



153 shared; 51 of 204 specific to each list

25.0%

Captured positives
Raw 36 Era 37

34 readmission and 51 ICU episodes differ in each list, alongside only 3 and 1 additional captured positives.

Priority-list variability under retraining

Mean Jaccard overlap of selected episode sets. A higher value means more shared membership.

Task	Seeds	Raw within	Era within	Cross representation
Readmission	1	0.475	0.546	0.530
Readmission	2	0.575	0.640	0.625
ICU transfer	1	0.289	0.370	0.342
ICU transfer	2	0.414	0.450	0.457

Within: disjoint seed subsets. Cross: matched seed indices. Pair comparisons are dependent.

Retraining itself changes priority lists. Cross-representation disagreement cannot be attributed uniquely to compression.

Evaluation choices can reverse point preferences

Each cell names the representation favored by the point summary. Color does not imply statistical significance.

Readmission

	Episode ensemble	Equal-patient ensemble	Seed majority
AUROC	Era	Era	Era
AP	Era	Era	Era
Brier	Era	Era	Era
Log loss	Era	Era	Era
P@10%	Era	Raw	Era

ICU transfer

	Episode ensemble	Equal-patient ensemble	Seed majority
AUROC	Era	Era	Raw
AP	Era	Era	Raw
Brier	Era	Era	Era
Log loss	Era	Era	Raw
P@10%	Era	Raw	Era

Equal-patient P@10 points to Raw in both tasks. All paired episode- and patient-weighted intervals include zero.

Calibration after tuning-only Platt scaling

Held-out ensembles, with 2,000 patient-cluster bootstrap samples. Targets: intercept 0 and slope 1.

Task	Pipeline	ECE	Intercept [95% CI]	Slope [95% CI]
Readmission	Raw	0.0226	0.354 [0.063, 0.602]	1.270 [1.119, 1.417]
Readmission	Era	0.0219	0.217 [-0.049, 0.446]	1.197 [1.061, 1.335]
ICU transfer	Raw	0.0128	0.563 [-0.129, 1.300]	1.217 [0.992, 1.454]
ICU transfer	Era	0.0118	0.439 [-0.238, 1.173]	1.178 [0.967, 1.431]

ECE uses 10 equal-frequency bins. Calibration quality and priority-list agreement measure different properties.

Both readmission slope intervals remain above 1. Tuning-set calibration does not ensure perfect held-out calibration.

ICU component controls

Era+backfill AP differences with nominal paired 95% CIs. Structure-null preserves length, times, base diagnoses and backfill but removes state tokens and values.

vs Raw

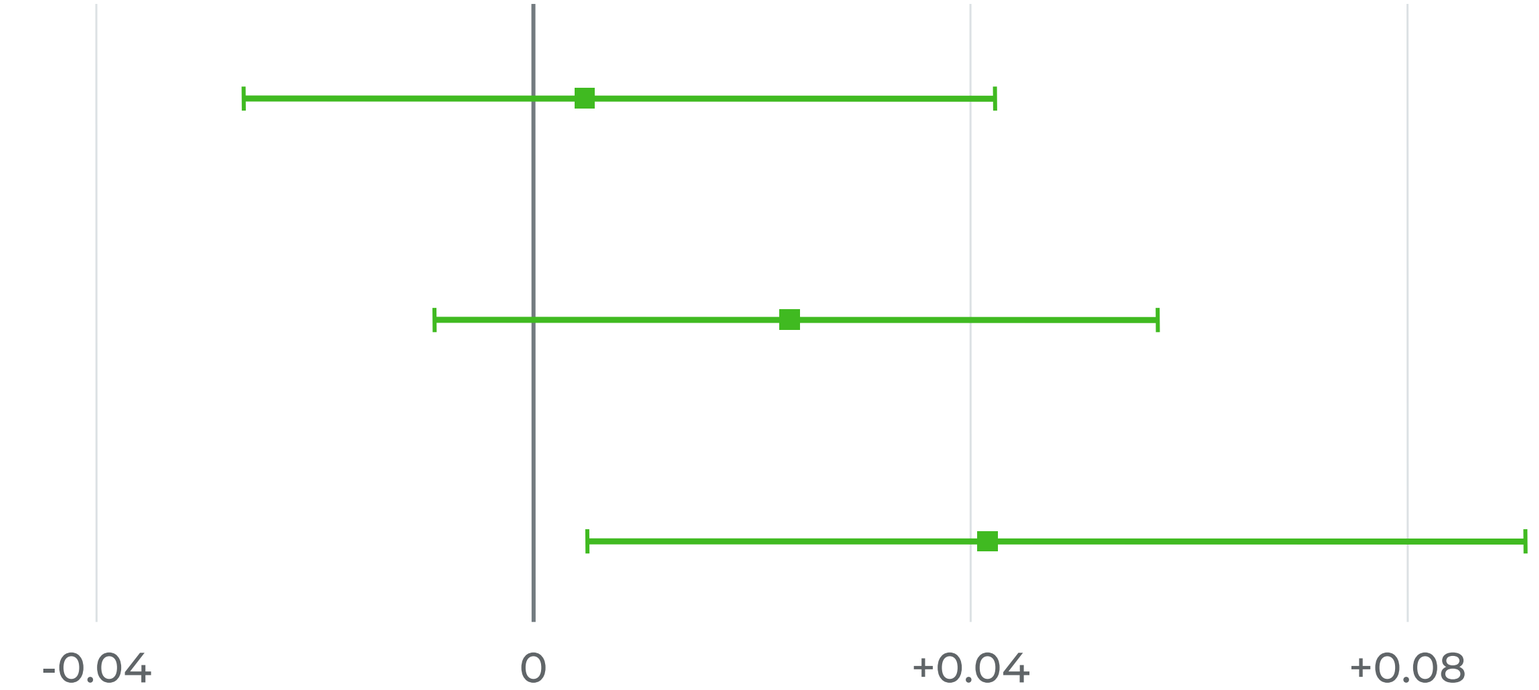
AP difference +0.0047

vs no-backfill

AP difference +0.0234

vs structure-null

AP difference +0.0416



No-backfill transforms the Raw suffix, changing both historical reach and era construction.

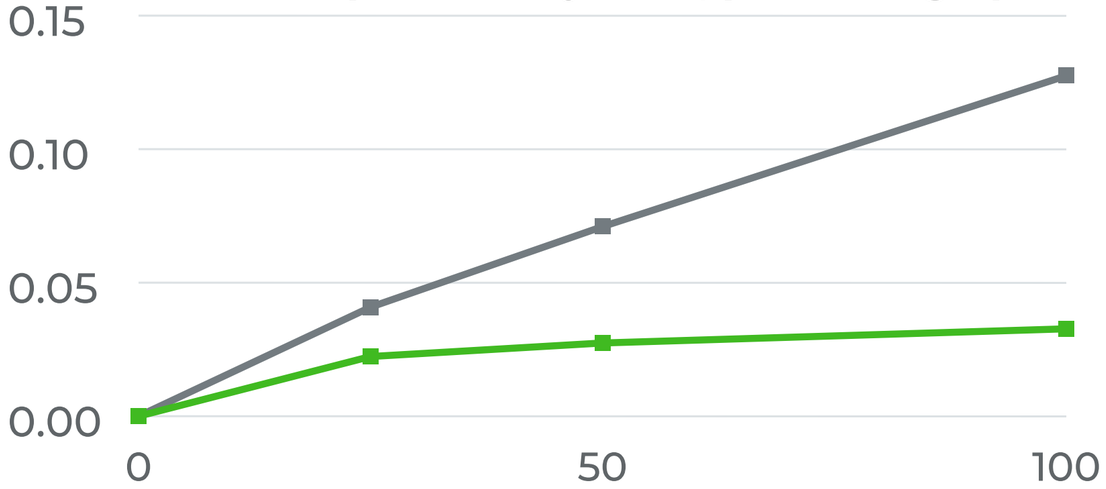
Retrospective sign-tail/Holm sensitivity across 60 quantities retains log loss only among five ICU state contrasts. No confirmatory testing.

Diagnosis-copying sensitivity reverses across tasks

Copy one persistent diagnosis seen in at least two visits into later visits where absent, strictly before prediction. Both pipelines receive identical insertions.

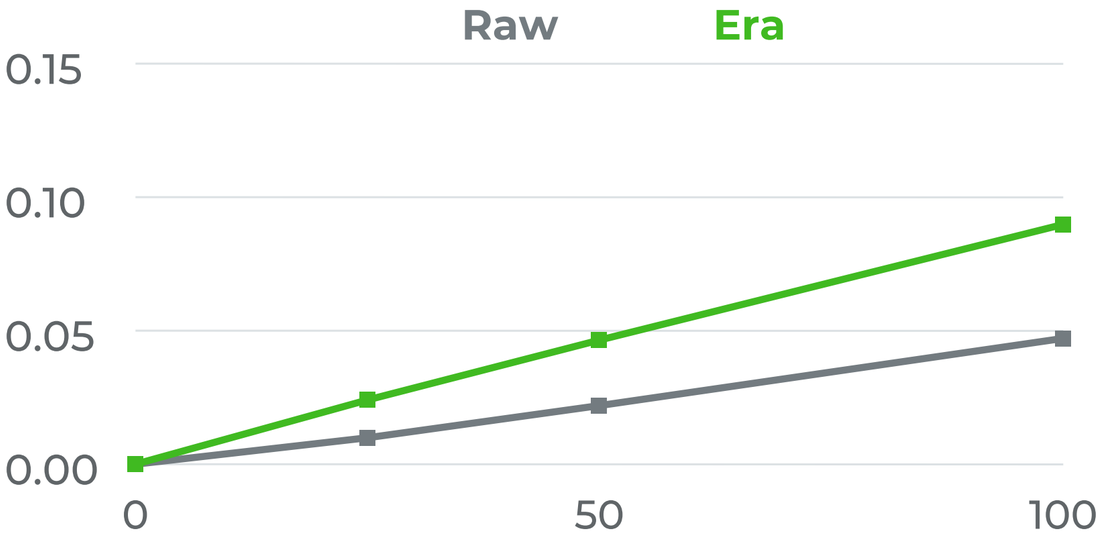
Readmission (n = 1,286)

Mean absolute probability shift, percentage points



Raw 0.128 pp Era 0.033 pp

ICU transfer (n = 1,028)



Raw 0.047 pp Era 0.090 pp

Requested fraction f (%). Insertions = ceil(f / 100 x eligible visits)

Synthetic copying: Era shifts 74% less than Raw for readmission and 91% more for ICU. Each uses its own 0% baseline.

Implications for clinical model selection

Aggregate quality, priority-list agreement and behavioral sensitivity provide distinct evidence.

15.5% / 25.0% of each top-decile list is representation-specific, despite close aggregate scores

Retraining also changes membership, so representation differences need like-sized references

Task-specific repetition sensitivity favors Era for readmission and Raw for ICU

One EHR source, 85 ICU positives and retrospective held-out reuse (81 variants/task). Next: a locked external cohort and a predeclared protocol.

Paper, code and contact

Public code includes the aggregate results used in this presentation.

Paper

OpenReview / RffArnWpec

Code

GitHub / LvovDmitro

Google Scholar

Dmitry Lvov

github.com/LvovDmitro/ehrshot-state-or-space

Record-level EHRSHOT data require authorized access. No patient records are distributed with the code.

Supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-10-2025-034 dd. 19.06.2025, IGK 000000C313925P4D0002).

Context length and era-horizon sensitivity

Five-seed calibrated ensembles. Longer context redistributes performance across metrics.

Task	Pipeline	L	Gap (days)	AUROC	AP	Brier	P@10%
Readmission	Raw	16,384	-	0.8034	0.4202	0.08804	0.4384
Readmission	Era	16,384	90	0.7942	0.4161	0.08767	0.4658
ICU transfer	Raw	16,384	-	0.7805	0.2017	0.03690	0.1814
ICU transfer	Era	16,384	90	0.7774	0.2352	0.03615	0.1765
ICU transfer	Era	4,096	30	0.7550	0.1601	0.03765	0.1569
ICU transfer	Era	4,096	90	0.8013	0.2241	0.03618	0.1814
ICU transfer	Era	4,096	180	0.7757	0.2362	0.03634	0.1961

Era = Era+backfill. Higher AUROC, AP and P@10 are better. Lower Brier is better.

Within tested settings, point-estimate trade-offs depend on task and metric. Era-gap sensitivity was evaluated only for ICU.