FROM FORECAST ERROR TO TRIAGE DECISIONS: PATIENT-CENTERED EVALUATION OF ICU VITAL-SIGN FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

In intensive care, a vital-sign forecast would be acted on through an early-warning score, a step function that bins each parameter into a few ordinal levels and sums the points. We therefore evaluate ICU vital-sign forecasts through that score. Each forecast is mapped through a four-parameter NEWS2-style risk index (respiratory rate, SpO_2, systolic blood pressure and heart rate), converted into triage levels and scored against two separate ground truths: the index on the observed vitals and events recorded in the 24 h after the forecast origin (termination, vasopressor start, dialysis start, invasive ventilation or any of these). On 347,648 windows from the MIMIC-IV dataset, the 9.3% masked-MSE lead of the best in-domain-trained model over the best zero-shot foundation models shrinks to 1% in weighted κ. With the same 9.6% of windows flagged in both states, so that every extra event recovered is one fewer false alarm, triage on a zero-shot TimesFM forecast recovers 18% more deterioration-bearing windows than triage on the currently observed vitals at H=6 h (sensitivity 12.4% to 14.7%; 95% CI on the gain 13.9 to 22.3, resampled over patients). Ranked by the raw subscore rather than by level, the gain is 9.0%; ranked by level, it decays to 7.2% at 24 h. Squared error does not select the forecast model family that recovers the most events: at a forced 10% budget, GRU, the model with the lowest MSE, recovers fewer deteriorations than every zero-shot model (Chronos-2 minus GRU +6.2 relative-gain points, 95% CI 3.0 to 9.4, paired on the same patients) and 27 of 36 paired zero-shot-minus-trained contrasts exclude zero. In addition, we report the harmonised benchmark behind the forecasts (MIMIC-IV-CLIF and eICU-CRD, six vitals, $L = 48$ h, $H = 1$ to 24 h, 16 models) and release the risk and triage protocols as secondary evaluations for vital-sign forecasting benchmarks.

1 INTRODUCTION

A vital-sign forecast reaches a clinical decision through a scoring rule of a different shape from the loss it was trained on. An early-warning score bins respiratory rate, oxygen saturation, blood pressure and pulse into four ordinal levels, from 0 to 3 points each and triage level is set on the total number (Royal College of Physicians, 2017). However, every vital-sign forecasting benchmark reports squared error, which is measured on a different loss function than that step function does. Specifically, a forecast model that wins a leaderboard might not when it comes to a risk index. We call this mapping, from a set of vital signs to a summed ordinal score and a triage level, the *read-out*.

In this work, we use the read-out as the evaluation. We compute the index on the forecast, ask whether the level it assigns anticipates what happens to the patient, and treat the squared-error leaderboard as the input to that evaluation rather than the result. Because the index is computed from vitals, agreement with the index on observed vitals would be reported. Two ground truths are therefore kept apart throughout (Section 5): agreement with the index recomputed on what was observed (a calibration statement) and discrimination of a recorded clinical event (the clinical claim). The recorded events break the circle and the two ground truths rank models differently.

Three results follow. First, the read-out compresses the leaderboard. Among the nine risk-scored models, the 9.3% best-to-best squared-error gap between in-domain training and zero-shot founda-

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.

tion forecasting becomes 1.0% in weighted κ and 0.7% in threshold AUROC, part of which is a quantisation ceiling of the index (Section A.3). Second, the read-out pays off in events recovered per alarm issued, the metric a triage policy is judged in. With 9.6% of windows flagged in both states, triage on a 6 h TimesFM forecast recovers 18.0% more deterioration-bearing windows than triage on the currently observed vitals (sensitivity 12.4% to 14.7%; 95% CI on the gain 13.9-22.3%), 9.0% if alarms are ranked by raw subscore rather than by level and the gain falls to 7.2% at $H{=}24$ h whilst the κ margin over the current vitals widens from $+0.1$ to $+0.14$. Third, squared error does not select the forecast model family that recovers the most events. At a 10% budget all three zero-shot models separate from the current vitals at 6 and 12h. None of the four trained models does and 27 of 36 paired zero-shot-minus-trained contrasts exclude zero, led by Chronos-2 over GRU, the squared-error winner ($+6.2$ relative-gain points, 95% CI 3.0-9.4).

Contributions:

1. **A risk-index read-out for vital-sign forecasts:** a four-parameter NEWS2-style subscore computed on the forecast and on the observed horizon mean, evaluated as an ordinal variable and turned into triage levels under three rate-calibrated schemes (Section 3).
2. **The leaderboard depends on the scoring rule:** the measurement that the leaderboard is a function of the scoring rule, and the family-level structure of that dependence (Section 4).
3. **Outcome-anchored triage:** forecast triage is scored against five recorded 24 h outcomes over $H \in \{6, 12, 24\}$ h, with patient-clustered intervals and three estimands kept apart (Sections 6 and 7).
4. **Squared error does not select the family:** at a forced 10% alarm budget, the model family with the lowest MSE is not the family that recovers the most events (Section 8).

The two-cohort benchmark behind the forecasts is the setting, not the finding (Sections A.1 and 2).

2 THE FORECASTING BASELINE

Data and task. Two adult critical-care cohorts are harmonised onto an hourly grid: MIMIC-IV through the CLIF schema (Johnson et al., 2023; Rojas et al., 2025) and eICU-CRD (Pollard et al., 2018). Six vitals are forecast jointly, heart rate, SpO_2, systolic and diastolic blood pressure, mean arterial pressure and respiratory rate, from an $L{=}48$ h look-back at horizons $H{=}1$ to 24 h, with patient-grouped $70/15/15$ splits. Every number below is on the held-out test split. Windows are cut at a 1 h stride over each hospitalisation's hourly record, so one clinical event is a positive for many overlapping windows, every count below is a window count with intervals clustered by patient and the origin lies inside an ICU episode for 83.4% of windows. Observation is irregular, so every metric is accumulated only over horizon cells that were actually observed (87.2% of cells) and forward-filling never crosses into a scored cell. The risk and triage analyses use the MIMIC cohort. The risk layer keeps complete-case windows from patients observed at least 95% complete with at least 20 such windows (347,648 windows over 2,223 patients). The triage layer keeps every complete-case window (542,987 windows of 4,403 patients at $H{=}6$ h, 460,391 at 12 h and 341,290 at 24 h).

Models and which of them are risk-scored. Sixteen models are evaluated: six networks trained in domain (GRU, LSTM, Mamba, Transformer, PatchTST, AR-NNet), six zero-shot Foundation Models (FM), TimesFM (Google Research, 2026), Chronos-Bolt (Ansari et al., 2024), Chronos-2 (Ansari et al., 2025) and TiRex (Auer et al., 2025) and four statistical baselines including last-observation persistence. A few-shot-adapted model is reported alongside them in Section A.1, outside the sixteen models. The trained models minimise masked squared error over the six vitals and are single-seed runs. The FMs are run univariate per vital in this study and the point forecast read from them is the mean head for TimesFM and the median quantile for Chronos-2 and TiRex. Nine of the sixteen are carried into the risk layer: the six trained networks and the three current-generation zero-shot models. The four statistical baselines and the three superseded checkpoints are not risk-scored, and persistence enters the risk layer as the currently observed vitals. The triage layer needs a per-window dump at every horizon, which seven of the nine carry: the three zero-shot models and GRU, LSTM, AR-NNet and PatchTST (Mamba and the Transformer were run at $H{=}6$ h only). In-domain training leads on squared error. Among the nine risk-scored models the best trained network

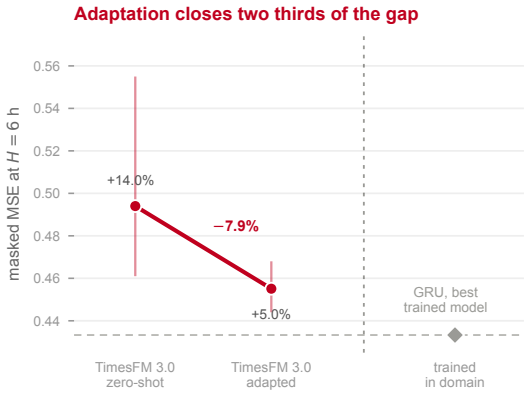


Figure 1: **What a small amount of in-domain data buys a foundation forecaster.** Masked MSE at $H{=}6$ h, MIMIC, all three points from Table 5. The connected version is one checkpoint before and after adaptation on 20,000 patient-grouped windows; bars are patient-clustered bootstrap intervals, whose narrowing is most of what adaptation achieves. The trained network past the divider is a different model and is left unconnected. Percentages are the distance above the dashed trained level.

(GRU, 0.4333) beats the best zero-shot model (TiRex, 0.4735) by 9.3% in masked MSE at $H{=}6$ h, and the gap widens with horizon from 6.4% at $H{=}1$ to 10.3% at $H{=}24$ in MIMIC, with eICU have similar results.

3 A NEWS2-STYLE RISK INDEX ON THE FORECASTED VALUES

The Risk Index. Six vitals are forecasted, four of them are NEWS2 parameters and those four are what the index reads: respiratory rate, SpO_2, systolic blood pressure and heart rate. Diastolic and mean arterial pressure are forecast and scored on the leaderboard but carry no NEWS2 points, so they enter the index only through their correlation with the four that do. Figure 2 walks one window through the read-out. Each of the four is assigned to its published ordinal levels worth 0-3 points and the points are summed to a subscore in 0-12, which is cut into four triage levels at the NEWS2 aggregate edges: T0=0, T1=1-4, T2=5-6, T3 $\geq$ 7 (Royal College of Physicians, 2017). The edges are applied unchanged to the 0-12 subscore rather than rescaled for the missing items, and the NEWS2 single-parameter trigger (any row scoring 3) and the SpO_2 Scale 2 for chronic hypercapnia are not implemented. The top level is therefore harder to reach than in NEWS2 proper, and identically so for every state. The index is a function of a *state*, so the same rule is applied to a forecast, to the vitals last observed at the forecast origin (the *currently observed vitals*, persistence or assume-no-change in forecasting terms) and to what was actually observed over the horizon (the *observed* state).

No recalibration before scoring. Each state is scored exactly as it comes out of the forecaster, no re-centring, variance inflation or threshold shift because the object of study is what the read-out does to a forecast as delivered, which is also why the under-call of Section 4 appears as a signed bias instead of being calibrated away. The ordinal map is lossy: exact subscore agreement runs 38.8-40.4% across the nine models and agreement to within one point 75.2-76.7%, so about 60% of all disagreement is a single-point boundary crossing. Section A.3 decomposes it.

Three schemes, rate-calibrated. We form triage levels three ways.

- *Population* uses the same fixed thresholds for every patient and therefore reads absolute severity.
- *Personalised* scores each vital against that patient's own 48 h baseline and therefore reads change.
- *Hybrid* takes, for each parameter, the higher of the two.

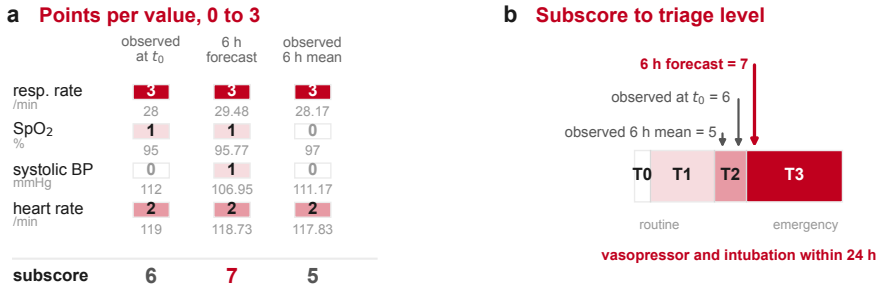


Figure 2: **Risk Index on a real patient.** **(a)** each state's value for the four scored vitals and the points it earns; **(b)** the three subscores on the 0-12 scale. The currently observed vitals score 6 and the 6 h forecast 7, moving this window from T2 to T3; vasopressor and invasive ventilation both started within 24 h. The observed horizon mean scores 5, *below* the forecast: agreement with the observed index and discrimination of the event are different questions (Section 5). The window was chosen by a rule fixed in advance (complete-case, forecast level above the current level under both schemes, a recorded deterioration: 1,744 windows, 0.32% of complete-case windows), then by closest-to-median look-back missingness. Figure 10 (Section A.3) draws this patient's vitals inside their bands.

Schemes are rate-calibrated on how often each parameter issues a point, not on how large a share of windows lands in each triage level, so level shares may be compared across states within one scheme but never across schemes. The risk layer of Section 4 uses the population bins, the NEWS2 scheme is as published, whilst the triage layer of Sections 6 to 8 uses the hybrid scheme. The three schemes are compared in Section A.1.

4 From Error to Risk

What changes when the same forecasts are scored as risk. Table 1 reports the nine risk-scored models under masked MSE and under the four-parameter index, on the same 347,648 windows from MIMIC dataset. The spread across models shrinks sharply: a gap of about 14% in masked MSE becomes about 1.6% in weighted κ. The best-to-best gap between the strongest in-domain model and the strongest zero-shot foundation model falls from 9.3% in MSE (GRU 0.4333 against TiRex 0.4735) to 1.0% in weighted κ and 0.7% in AUROC for an observed subscore of at least 5. The pair carried into triage, AR-NNet against TimesFM, still differs by 11.8% in MSE (0.44 against 0.49), but that gap is no longer the story once the index is the metric.

The ranking does not reverse in full. The families interleave: TimesFM is last of the nine on squared error and sixth on κ, above one trained network. Spearman correlation between masked MSE and weighted κ is $\rho = -0.77$ ($p = 0.016$), so the fair reading is compression of the gaps with some local reordering, not a complete swap of winners. Figure 6 shows the same contrast.

Every forecast under-calls risk. Signed bias in subscore points runs from -0.29 to -0.11 across the nine forecasts. Naive persistence (the index on the currently observed state) is the only row biased the other way ($+0.28$). Within that range, the pattern is consistent: the in-domain models, which win on MSE, under-call the most (about -0.24 to -0.29) and the zero-shot models under-call the least (about -0.11 to -0.17). Squared error rewards a forecast that stays near the middle of the range. A Risk Index then scores that same forecast as low level, zero points. We leave that bias in place on purpose: recalibrating before scoring would hide the effect we are measuring.

5 Deterioration as Ground Truth

The index is computed from vitals, so agreement with the index recomputed on observed vitals cannot by itself establish clinical value. We therefore score every state against two stages.

Table 1: The same forecasts under squared error and under the four-parameter risk index, on 347,648 MIMIC windows over 2,223 patients. Weighted κ and AUROC are against the subscore of the observed horizon mean; bias is the signed mean difference in subscore points, negative meaning the forecast under-calls risk. The persistence row is the index read on the currently observed state.

Model	masked MSE	weighted κ	AUROC (≥ 5)	bias
trained in domain				
GRU	0.4333	0.7476	0.8890	-0.28
LSTM	0.4337	0.7539	0.8909	-0.24
Mamba	0.4347	0.7494	0.8887	-0.27
Transformer	0.4349	0.7506	0.8892	-0.27
PatchTST	0.4394	0.7464	0.8870	-0.28
AR-NNet	0.4418	0.7446	0.8865	-0.29
foundation, zero-shot				
TiRex	0.4735	0.7447	0.8840	-0.14
Chronos-2	0.4799	0.7419	0.8825	-0.11
TimesFM	0.4940	0.7462	0.8845	-0.17
persistence (current state)	0.6792	0.6511	0.8491	0.28

Stage 1: Risk Index calibration. Stage 1 is agreement with the subscore of the *observed* horizon mean, a function of the same six vitals. It is a calibration statement, and the observed state has $\kappa = 1$ by construction. It is a ceiling, not a clinical claim.

Stage 2: Recorded outcomes. Stage 2 is discrimination of *recorded events* in the following 24 h: death or discharge to hospice ("termination" hereafter), first vasopressor start (norepinephrine, epinephrine, phenylephrine or vasopressin), first dialysis start (continuous renal replacement therapy; intermittent haemodialysis is not captured), first invasive ventilation ("intubation" hereafter) and their union, the *termination* or "any deterioration" (61,916 event-bearing windows). "Deterioration" below always means this treatment-escalation termination. None of the five is computable from the vitals themselves, and the observed state, a ceiling in Stage 1, is beaten by the 6 h forecast on two of them in Stage 2 (Section A.1). Section A.5 tests the read-out on remaining ICU stay.

The Outcome. For a window whose look-back ends at the forecast origin t_0, the outcome interval is $[t_0,\, t_0 + 24\,\mathrm{h})$. It opens at the origin, not at $t_0{+}H$, and it is the *same* absolute window at every forecast horizon, so only the state the index reads changes with H: the outcome is fixed. This is what makes the horizons of Section 7 comparable. Because a window exists only if its horizon cells were observed, a positive is in practice an event in $[t_0{+}H,\, t_0 + 24\,\mathrm{h})$ among patients monitored through $t_0{+}H$ (Section 7 gives the consequence for prevalence). Onsets are episode-based and each is restricted to windows where it was not already running at t_0 (Section A.3).

6 Triage at a fixed alarm budget

Having separated Stage 1 from Stage 2, we turn the same four triage levels into an alarm and ask whether a forecast state recovers more recorded deteriorations than the currently observed vitals when both are held to the same review budget.

At six hours under the hybrid scheme, triage on the forecast beats triage on the currently observed vitals on all five Stage 2 outcomes (terminations $0.61 \to 0.65$, invasive ventilation $0.53 \to 0.58$, termination $0.56 \to 0.58$ in AUROC. Table 6 in Section A.1). Absolute discrimination remains modest: every scheme $\times$ state $\times$ outcome cell lies between 0.52 and 0.67, so these numbers support a comparative ordering of read-outs rather than a deployable alarm. The top level concentrates rather than inflating. Under currently observed vitals, T3 holds 20% of windows at a 14.7% deterioration rate, under the forecast it holds 9.6% at 17.4% and the top-to-bottom risk ratio rises from 1.69 to 2.29 (observed-state ceiling 2.52).

Because the two states flag different fractions of windows, a fixed level threshold is not a fair comparison: the forecast would be improving precision with recall and F_1 could move either way. We therefore match the alarm budget. Flagging 9.6% of windows under both states, triage on the fore-

cast recovers 9,094 of 61,916 deteriorations against 7,704 on the currently observed vitals: 18.0% more events for the same number of alarms, with precision 14.7% → 17.4%, F_1 0.135 → 0.159 and lift over base rate 1.29 → 1.52. Terminations gain most (47.9%: 1,964 against 1,328) and every outcome gains (Figure 3c). Two caveats belong with the gain. Operating points come from per-level event rates with uniform within-level allocation, which assumes a flat event rate inside an ordinal level and an F_1 of 0.16 at 11.4% prevalence is what a four-parameter index supports.

Table 2: The Risk Index at each forecast horizon, one model throughout (zero-shot TimesFM). Agreement is weighted κ against the subscore of the observed horizon mean. The right-hand block is the hybrid scheme on 24 h any deterioration, current vitals → forecast, at a matched 9.6% alarm budget, level-ranked. *Scored windows* are risk-layer counts; the triage block uses 542,987, 460,391 and 341,290 windows.

	risk index			*triage, matched budget*			
H (h)	scored windows	κ	margin	events	recall	precision	gain
6	347,648	0.746	+0.095	61,916	12.4% → 14.7%	14.7% → 17.4%	+18.0%
12	298,839	0.740	+0.122	52,610	12.2% → 13.9%	14.4% → 16.5%	+14.2%
24	228,204	0.719	+0.135	39,169	11.7% → 12.6%	14.0% → 15.0%	+7.2%

The 9.6% budget is the share of windows that the TimesFM forecast itself places in T3, and the currently observed vitals are held to the same count. This choice favours the forecast: its alarm set is exactly its own T3 level, whereas the T3 of the currently observed vitals (20% of windows) is cut to the budget at random within the level. Ranking alarms by the raw subscore removes that advantage and halves the gain to 9.0%. Because both states issue the same number of alarms over the same windows, every extra true positive is one fewer false positive. The read-out therefore reorders which windows are flagged at a fixed review capacity rather than trading alarm burden for sensitivity; the matched counts and specificity are in Section A.2.

The interval analysis forces the budget to exactly 10% on the 341,290 windows common to all three horizons (39,169 events over 3,046 patients), resampling patients 1,000 times (Section A.2). The forecasted values volume is smaller there because both the forced budget and the common window set cost effect size. Two parameter-free smoothing controls separate the gain from look-back averaging: the same index scored on the mean of the last 6 h of observed values gains 1.1% (−2.2 to 4.5%) and on the last 24 h −1% (−6.5 to 4.5%), both intervals covering zero, against 9.4% (5.7-13.1%) for the observed-state comparator. The gain therefore comes from the forecasted future state, not from smoothing the recent past.

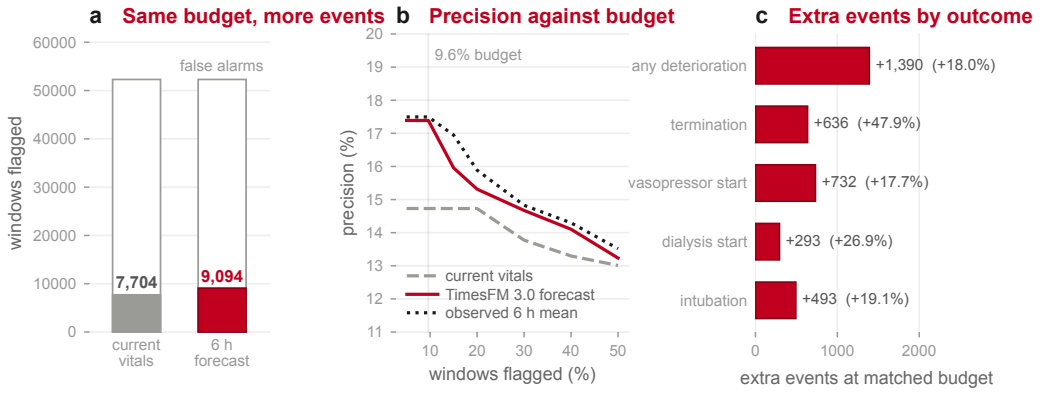


Figure 3: The gain in the currency of alarms, composite outcome under the hybrid scheme at 6 h. **(a)** both models flag the same 52,300 windows, and the shaded part is the deteriorations among them; **(b)** precision against alarm budget, the flat segments being the within-level uniform allocation; **(c)** extra events recovered at the matched budget, by outcome.

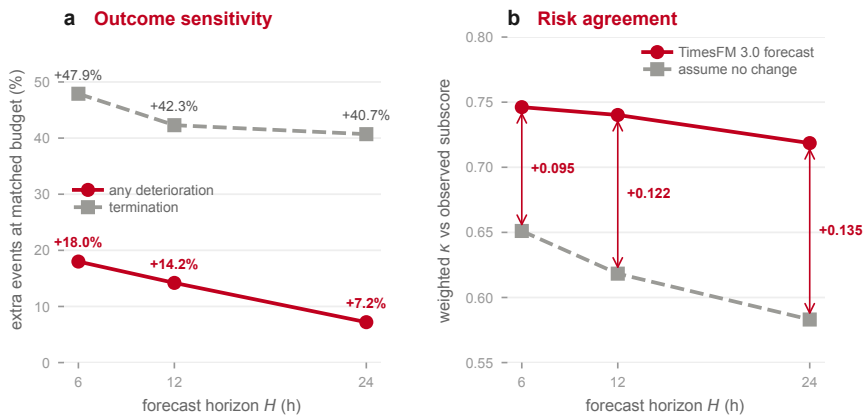


Figure 4: Stage 1 and Stage 2 as the horizon grows, one model throughout (zero-shot TimesFM). **(a)** extra events recovered over triage on the currently observed vitals at the matched 9.6% budget, level-ranked: the composite gain falls to 40% of its 6 h value. The termination line is read within a horizon only, because its eligible population changes with H (Section 7). **(b)** weighted κ against the observed subscore: the forecast decays slowly while the current vitals fall away, so the margin widens.

7 Across forecast horizons

The read-out repeats at any horizon the forecasting baseline supports. We therefore rerun the forecast on one model, zero-shot TimesFM, at H=6, 12 and 24 h and change nothing else, down to the same absolute 24 h outcome window. Among the three zero-shot models it has the smallest forced-budget gains at 12 and 24 h (Table 4), so the gains below are the most conservative of the three and its decay with horizon is the steepest.

Stage 1 and Stage 2 move in opposite directions. Against the currently observed vitals, the Stage 1 margin on the risk index widens monotonically with horizon: +0.095, +0.122, +0.135 in weighted κ at H=6, 12, 24 h (Figure 4b), mostly because the agreement of the current vitals decays. The level-ranked Stage 2 sensitivity gain over the same baseline narrows over exactly the same range: +18.0% (13.9-22.3%), +14.2% (9.7-18.5%) and +7.2% (1.7-12.7%) extra deteriorations at the matched 9.6% budget (Figure 4a; Table 2 in Section A.2). This is consistent both with the event being harder to anticipate from a longer forecast and with a longer horizon mean diluting the excursions that carry the signal (Section A.4), and the design does not separate the two. Ranked by subscore, the 12 and 24 h gains are 7.3% (3.2-11.3%) and 3.0% (−1.9 to 8.0%), so the 24 h gain is not robust to the ranking rule. The index-agreement margin is therefore largest where squared error is worst, but only the outcome-anchored stage speaks to utility, and its case is strongest at short horizons.

What is comparable across horizons and what is not. The alarm budget is matched at 9.6% of windows at every horizon, and the composite keeps essentially the same prevalence (11.4, 11.4, 11.5%), so precision and lift are directly comparable. Two quantities are not. First, termination prevalence falls from 1.5% at H=6 h to 0.2% at H=24 h, because a window needs observed future cells to exist and a patient who has died has none. The onset outcomes inherit the same mechanism through their masks, so termination and onset gains are read within a horizon and never as a trend. Second, the window population differs by horizon (Section 2), so the interval analysis below is run on the intersection.

The decay is real for this model. On the common window set at a forced 10% budget (Section A.2), the same ordering carries intervals. TimesFM recovers 6.2% more deteriorations than the currently observed vitals at H=6 h (95% CI 1.4-10.7%), 5.8% at 12 h (1.1-10.6%) and 3.7% at 24 h (−1.4 to 8.9%), the last interval covering zero. For this model the decay is separated from zero: the H=24 minus H=6 difference in recall is −0.31 percentage points (−0.55 to −0.03), and no other

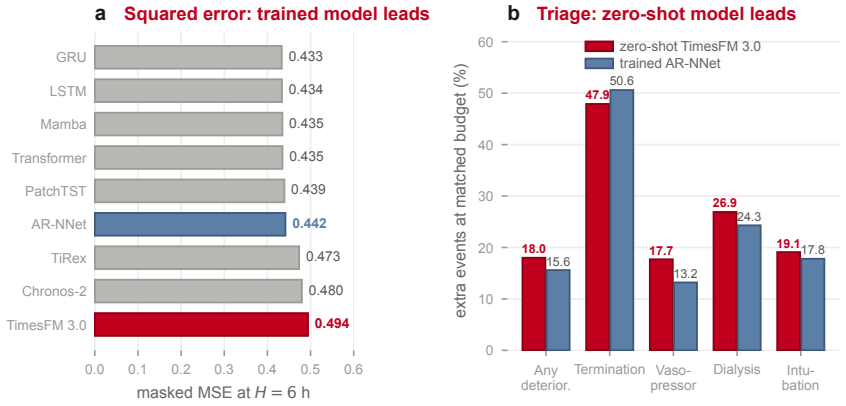


Figure 5: Squared error and triage rank the two headline models differently, $H{=}6$ h; point estimates, and the paired contrast between these two models is not resolved (Section 8). **(a)** masked MSE over the nine risk-scored models, with the two triage models highlighted; **(b)** extra events recovered at the level-ranked 9.6% alarm budget, by outcome. The contrast that does separate, GRU against Chronos-2, is in Table 3.

model shows a separated decay (Section A.2). For TimesFM, forecast-driven triage therefore shows evidence of benefit at 6 and 12 h, whilst its 24 h result is inconclusive. Chronos-2 and TiRex still separate at 24 h (Table 4).

8 Which forecast model.

The triage stage at $H{=}6$ h carries one trained model, AR-NNet, alongside zero-shot TimesFM and the two disagree in a way that matters for what a benchmark should report. On squared error AR-NNet leads by 11.8% (0.44 against 0.49 masked MSE, Figure 5a), and it is ahead on triage AUROC on all five outcomes by 0.001-0.002. That AUROC ordering does not survive a paired test: on the common window set the contrast is -0.0004 (95% CI -0.0038 to $+0.0029$), covering zero and nominally favouring the foundation models instead. At the matched alarm budget the ordering reverses on four of the five outcomes: TimesFM recovers $+18\%$ against $+15.6\%$ on the composite, $+17.7$ against $+13.2\%$ on vasopressor start, $+26.9$ against $+24.3\%$ on dialysis start and $+19.1$ against $+17.8\%$ on intubation, with AR-NNet ahead on termination alone ($+50.6$ against $+47.9\%$; Figure 5b). On risk agreement TimesFM also leads, by a small margin ($+0.0016$) in κ.

The model family-level effect. A single pair of models would be an anecdote, so Section A.2 runs the budget-matched comparison for all seven models at all three horizons with patient-clustered intervals. At the 10% budget the split falls by family rather than by model. All three zero-shot models separate from the currently observed vitals at $H{=}6$ and 12 h (TimesFM $+6.2\%$, Chronos-2 $+8\%$, TiRex $+7\%$ at $H{=}6$ h, every interval excluding zero) and two of the three still do at $H{=}24$ h. None of the four trained models separates at any of the three, GRU $+1.9\%$, LSTM $+2.5\%$, PatchTST $+2.5\%$, AR-NNet $+3.5\%$ at $H{=}6$ h. Chronos-2 is the strongest model under this read-out at every horizon.

A paired test separates the model family. Contrasting models on the same resamples, rather than each against the currently observed vitals, splits the claim in two. Across all 36 zero-shot-versus-trained contrasts, three zero-shot models, four trained models, three horizons, 27 put the zero-shot model ahead at the matched budget with an interval excluding zero, but they are unevenly distributed: Chronos-2 separates in 12 of 12, leading GRU by 6.2 relative-gain points at $H{=}6$ h (95% CI 3.0-9.4) and AR-NNet by 4.6 (1.7-7.5), TiRex in 10 of 12 and TimesFM in only 5 of 12 (Table 3). The specific pair this section opened with is one of the unresolved ones: TimesFM minus AR-NNet on the composite is $+2.7$ relative-gain points, 95% CI -0.2 to $+5.2$, covering zero at every budget from 5 to 20% and on all five outcomes separately. So the 18.0 against 15.6% of

Figure 5b is a point-estimate difference this design cannot resolve, while the group-level statement is resolved. The defensible reading is the weaker and more general one: squared error does not select the forecast model that recovers the most events at a fixed review capacity and it is the zero-shot models rather than any one checkpoint that carries the effect.

9 LIMITATIONS

The Risk Index is a four-parameter alternative to NEWS2, not NEWS2 itself, and about half of the scored windows are under organ support at t_0 (Section 3). Absolute discrimination is modest across all sixty triage cells (0.518-0.667), so this is a comparative result and not an early-warning score fit for deployment (Gerry et al., 2020). The headline gain depends on the ranking rule (18.0% by level against 9.0% by subscore) and on the binning scheme (the population-scheme matched-budget gain was not computed). The family split of Section 8 is resolved at the budget only and not for the headline pair. Because every in-domain model is trained under squared error and every zero-shot model under a different objective, the family split cannot be separated from a loss-level split. A trained model fitted under a quantile loss, or recalibrated with the level slope of Section A.4, is the control this paper lacks. The trained models are also single-seed runs.

Three coverage limits bound how far the result travels. The risk and triage results are results from MIMIC dataset: both cohorts carry the leaderboard and rank the families identically on it (Figure 6a), but the per-window dumps and label arrays exist only for the MIMIC split. Both subsets are sampled hourly and a binned index is sensitive to sampling density (Section A.4), so repeating the read-out on HiRID (Yèche et al., 2021), AmsterdamUMCdb (Thoral et al., 2021) and SICdb (Rodemund et al., 2024) is the test that would separate a property of the read-out from a property of these two databases, and is part of our future work. And the baseline holds no architecture built for sparse cross-channel ICU dynamics (Chang et al., 2024) and no multivariate foundation model run multivariately (Woo et al., 2024), so it is not a ceiling on what is predictable here.

10 CONCLUSION

This paper scores ICU vital-sign forecasts the way they would be acted on, through an early-warning score, rather than by squared error alone. Forecasts from in-domain trained networks and zero-shot foundation models are mapped through a four-parameter NEWS2-style Risk Index and scored against two ground truths: the index on the observed vitals and events recorded in the following 24 h. Three results follow. **The read-out compresses the leaderboard:** the 9.3% best-to-best squared-error gap between the two families is 1% in weighted κ. **Forecasting pays off in events recovered per alarm:** with 9.6% of windows flagged in both states, triage on a 6 h TimesFM forecast recovers 18.0% more deteriorations than triage on the currently observed vitals, 9.0% when alarms are ranked by subscore, and the gain falls to 7.2% at $H{=}24$ h. **Squared error does not select the model family that recovers the most events:** at a 10% budget GRU, the squared-error winner, is beaten by every zero-shot model at 6 and 12 h, and 27 of 36 paired zero-shot-minus-trained contrasts exclude zero. The family split cannot yet be separated from the training loss (Section 9), but the practical point does not depend on it. On squared error the zero-shot models trail the trained networks, and adapting TimesFM on $20{,}000$ in-domain windows closes about two thirds of its gap to the best of them. Under the read-out, the same zero-shot models are competitive with no training on this cohort at all. We therefore release the Risk Index and triage protocols as secondary evaluations to report alongside squared error on any vital-sign forecasting benchmark.

REPRODUCIBILITY STATEMENT

Code, configuration and evaluation scripts for this paper will be released in a public GitHub repository upon acceptance. MIMIC-IV and eICU-CRD are available to credentialed users under their respective data use agreements; the harmonisation follows the public CLIF schema (Rojas et al., 2025). The repository includes the risk and triage code with the three indexing schemes. Raw EHR extracts cannot be redistributed; credentialed users can rebuild the forecast store from the cited PhysioNet releases under the same protocol.

AI USE STATEMENT

In this work we used AI tools (Claude Opus 5 and Grok 4.6) for brainstorming, planning and coding: exploring evaluation ideas, outlining experiments, paper structure and checking that quoted numbers matched the saved result tables. We have not used generative AI tools to generate data, results or citations. We reviewed all AI-assisted text and code; LLM-assisted code was tested for correctness against the saved result tables and every quoted number and reference was verified by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work is a retrospective secondary analysis of publicly available, de-identified critical-care records hosted on PhysioNet (Goldberger et al., 2000): MIMIC-IV (Johnson et al., 2023) and the eICU Collaborative Research Database (Pollard et al., 2018). Access was obtained under the respective PhysioNet credentialed-user data use agreements. There was no intervention and no patient contact, and no attempt was made to re-identify individuals. Harmonisation onto the hourly grid follows the public CLIF schema (Rojas et al., 2025). The NEWS2-style risk index and triage read-outs reported here are research evaluations of forecasting practice; they are not validated for clinical use and are not proposed as deployable early-warning scores. The claims are comparative statements about how forecasts are scored, not recommendations for bedside decision support.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From Univariate to Universal Forecasting, 2025. URL https://arxiv.org/abs/2510.15821. arXiv:2510.15821.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-Shot Forecasting Across Long and Short Horizons with Enhanced In-Context Learning. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-1929. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5356603f9c47399adfd372f77a677057-Abstract-Conference.html.

Ping Chang, Huayu Li, Stuart F. Quan, Shuyang Lu, Shu-Fen Wung, Janet Roveda, and Ao Li. A transformer-based diffusion probabilistic model for heart rate and blood pressure forecasting in intensive care unit. *Computer Methods and Programs in Biomedicine*, 246:108060, 2024. doi: 10.1016/j.cmpb.2024.108060.

Stephen Gerry, Timothy Bonnici, Jacqueline Birks, Shona Kirtley, Pradeep S. Virdee, Peter J. Watkinson, and Gary S. Collins. Early warning scores for detecting deterioration in adult hospital

patients: systematic review and critical appraisal of methodology. *BMJ*, 369:m1501, 2020. doi: 10.1136/bmj.m1501.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Google Research. TimesFM 3.0. https://github.com/google-research/timesfm, 2026. Model checkpoint and code, release v3.0.0 (28 August 2026). Accessed 25 September 2026.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x. URL https://doi.org/10.1038/s41597-022-01899-x.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, 2018. doi: 10.1038/sdata.2018.178. URL https://doi.org/10.1038/sdata.2018.178.

Niklas Rodemund, Andreas Kokoefer, Bernhard Wernly, and Crispiana Cozowicz. Salzburg Intensive Care database (SICdb), a freely accessible intensive care database. *PhysioNet*, September 2024. doi: 10.13026/8m72-6j83. URL https://doi.org/10.13026/8m72-6j83. Version 1.0.8.

Juan C. Rojas, Patrick G. Lyons, Kaveri Chhikara, Vaishvik Chaudhari, Sivasubramanium V. Bhavani, Muna Nour, Kevin G. Buell, Kevin D. Smith, Catherine A. Gao, Saki Amagai, Chengsheng Mao, Yuan Luo, Anna K. Barker, Mark Nuppnau, Michael Hermsen, Jay L. Koyner, Haley Beck, Rachel Baccile, Zewei Liao, Kyle A. Carey, Brenna Park-Egan, Xuan Han, Alexander C. Ortiz, Benjamin E. Schmid, Gary E. Weissman, Chad H. Hochberg, Nicholas E. Ingraham, and William F. Parker. A common longitudinal intensive care unit data format (CLIF) for critical illness research. *Intensive Care Medicine*, 51(3):556–569, 2025. doi: 10.1007/s00134-025-07848-7. URL https://doi.org/10.1007/s00134-025-07848-7.

Royal College of Physicians. National early warning score (news) 2: Standardising the assessment of acute-illness severity in the nhs. updated report of a working party. Technical report, Royal College of Physicians, London, 2017. URL https://www.rcp.ac.uk/resources/national-early-warning-score-news-2/. Working party report, 19 December 2017.

Patrick J. Thoral, Jan M. Peppink, Ronald H. Driessen, et al. Sharing ICU patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: The amsterdam university medical centers database (AmsterdamUMCdb) example. *Critical Care Medicine*, 49(6):e563–e577, 2021. doi: 10.1097/CCM.0000000000004916.

Robin van de Water, Hendrik Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet Another ICU Benchmark: A Flexible Multi-Center Framework for Clinical ML. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ox2ATRM90I.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified Training of Universal Time Series Forecasting Transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. Hirid-icu-benchmark–a comprehensive machine learning benchmark on high-resolution icu data. *arXiv preprint arXiv:2111.08536*, 2021.

A Appendix

A.1 The forecasting benchmark and Risk Index

Table 5 is the 16-model leaderboard at $H{=}6$ h with skill against Naive persistence. Over all sixteen models the best trained network beats the best zero-shot model by 7.8% in masked MSE (GRU 0.4333 against TimesFM 0.4672). The 9.3% of the main text is the same gap among the nine risk-scored models, which exclude TimesFM, so it is taken against TiRex. The adapted row is TimesFM fine-tuned on this paper's protocol (20,000 patient-grouped windows, three epochs, learning rate 5×10^{-5}, AdamW, backbone unfrozen, univariate per vital, loss masked to observed horizon cells): masked MSE 0.4940 $\rightarrow$ 0.4551, a paired improvement of 7.9% (95% CI 3.6-15.5%, patient-clustered) on identical windows. What adaptation earns here is mostly the removal of a tail: the zero-shot bootstrap interval is 0.461-0.555 against 0.444-0.468 adapted, 3.9 times narrower.

Against the best trained model in the same table (GRU, 0.4333), the zero-shot gap is +14.0% and the adapted gap +5.0%: 20,000 windows of in-domain data close 64% of the distance and a model that never saw this cohort in pre-training ends within five per cent of a network trained on it (Figure 1). The residual 5.0% is what not training in domain costs. Three things to keep in mind. It is one model on one protocol at one horizon, as Section 9 states. The compared model is the best of six trained models rather than a fixed one. And because the adapted point is the same checkpoint before and after, this is a statement about adaptation and not about a newer release of TimesFM.



Figure 6: **(a)** within-dataset masked MSE at 6 h, MIMIC filled and eICU open; **(b)** the same nine models under masked MSE against weighted κ. The two panels use the two masked-MSE conventions present in the source tables and are not comparable across panels; each comparison is within one convention.

Figure 7 is the two-ground-truth comparison of Section 5, Figure 8 compares the three binning schemes and the four states, Figure 9 aggregates the risk read-out to the individual patient and Figure 11 carries the mechanism and the sensing-density ablation.

The two stages are different measurements and the clearest sign of it is that the observed-state comparator changes role between them. It is a ceiling in Stage 1 and an ordinary comparator in Stage 2, where the 6 h forecast *exceeds* it on two of the five outcomes, dialysis start and intubation (Figure 7). The observed state here is the realised 6 h horizon mean, not a reading of the present, and why a forecast of that mean should discriminate these two events better than its realised value is not established by this analysis.

Two limits come from the data itself. Neither cohort records a triage decision, a NEWS2 score or an acuity band, so a constructed index can be validated against events and never against what clinicians did. Minute-scale definitions are also unreachable on a 60-minute grid: sepsis has no

Table 3: Zero-shot minus trained, paired on the same resamples: the difference in composite deteriorations recovered at the forced 10% budget, in relative-gain points against the currently observed vitals, so +6.2 means the zero-shot model's gain over the currently observed vitals is 6.2 points larger than the trained model's. Positive favours the zero-shot FM. * marks a 95% interval that excludes zero. Chronos-2 separates from every trained model at every horizon.

zero-shot model	**trained model**	H=6 h	H=12 h	H=24 h
Chronos-2	GRU	$+6.2\ [+3.0, +9.4]^*$	$+6.7\ [+3.0, +10.7]^*$	$+7.6\ [+3.2, +12.1]^*$
	LSTM	$+5.5\ [+2.2, +8.8]^*$	$+7.9\ [+4.1, +11.9]^*$	$+6.9\ [+2.6, +11.1]^*$
	AR-NNet	$+4.6\ [+1.7, +7.5]^*$	$+5.0\ [+1.7, +8.3]^*$	$+4.8\ [+0.8, +8.9]^*$
	PatchTST	$+5.5\ [+2.6, +8.4]^*$	$+5.0\ [+1.6, +8.2]^*$	$+5.0\ [+0.7, +9.1]^*$
TimesFM	GRU	$+4.3\ [+1.1, +7.1]^*$	$+4.3\ [+0.8, +7.6]^*$	$+3.9\ [-0.3, +8.0]$
	LSTM	$+3.7\ [+0.4, +6.6]^*$	$+5.5\ [+1.6, +9.1]^*$	$+3.2\ [-0.9, +7.2]$
	AR-NNet	$+2.7\ [-0.2, +5.2]$	$+2.6\ [-0.5, +5.7]$	$+1.0\ [-2.4, +4.7]$
	PatchTST	$+3.7\ [+0.8, +6.2]^*$	$+2.6\ [-0.5, +5.5]$	$+1.2\ [-2.6, +4.8]$
TiRex	GRU	$+5.1\ [+2.2, +8.0]^*$	$+5.5\ [+2.1, +9.1]^*$	$+5.9\ [+2.0, +9.9]^*$
	LSTM	$+4.5\ [+1.2, +7.5]^*$	$+6.7\ [+3.2, +10.4]^*$	$+5.2\ [+1.2, +8.9]^*$
	AR-NNet	$+3.5\ [+1.0, +6.1]^*$	$+3.8\ [+0.6, +6.9]^*$	$+3.1\ [-0.5, +6.3]$
	PatchTST	$+4.5\ [+1.8, +7.1]^*$	$+3.7\ [+0.6, +6.9]^*$	$+3.2\ [-0.5, +6.7]$

Table 4: Budget-matched gain over triage on the current vitals for every model with a per-window dump, at a forced 10% alarm budget on the common window set. *gain* is the relative increase in combined deteriorations recovered, *spec.* the specificity at that budget in per cent. Intervals are percentile 95% from 1,000 patient-level resamples; * marks an interval excluding zero. The four trained models at H=12 and H=24 h were trained for this analysis on the same protocol as the H=6 h models.

	H=6 h			H=12 h			H=24 h		
Model	gain	95% CI	spec.	gain	95% CI	spec.	gain	95% CI	spec.
foundation, zero-shot									
TimesFM	$+6.2^*$	$[1.4, 10.7]$	90.43	$+5.8^*$	$[1.1, 10.6]$	90.43	$+3.7$	$[-1.4, 8.9]$	90.39
Chronos-2	$+8.0^*$	$[3.7, 12.5]$	90.46	$+8.2^*$	$[3.5, 13.1]$	90.47	$+7.5^*$	$[2.2, 13.0]$	90.46
TiRex	$+7.0^*$	$[2.9, 11.2]$	90.45	$+7.0^*$	$[2.5, 11.9]$	90.45	$+5.8^*$	$[0.9, 10.8]$	90.43
trained in domain									
GRU	$+1.9$	$[-2.3, 6.5]$	90.36	$+1.5$	$[-2.8, 6.5]$	90.36	-0.1	$[-5.2, 5.4]$	90.33
LSTM	$+2.5$	$[-1.8, 7.1]$	90.37	$+0.3$	$[-3.8, 4.9]$	90.34	$+0.6$	$[-4.1, 6.1]$	90.34
AR-NNet	$+3.5$	$[-0.8, 8.0]$	90.39	$+3.2$	$[-1.1, 7.9]$	90.38	$+2.7$	$[-1.9, 8.1]$	90.38
PatchTST	$+2.5$	$[-1.8, 7.1]$	90.37	$+3.2$	$[-1.0, 8.1]$	90.39	$+2.6$	$[-2.2, 8.1]$	90.37
current vital	recall 12.57%, specificity 90.33%, precision 14.42%: the reference for every row								

window-aligned label array on this split and a saturation below threshold sustained for ten minutes cannot be expressed.

Table 5: Masked MSE at $H=6$ h on the harmonised benchmark (window-mean convention) and skill against last-observation persistence. Lower MSE is better, higher skill is better.

Model	**masked MSE**	**skill vs. persistence (%)**
trained in domain		
GRU	0.4333	36.2
LSTM	0.4337	36.1
Mamba	0.4347	36.0
Transformer	0.4349	36.0
PatchTST	0.4394	35.3
AR-NNet	0.4418	34.9
foundation, adapted (this work)		
TimesFM (adapted)	0.4551	33.0
foundation, zero-shot		
TimesFM 2.5	0.4672	31.2
Chronos-Bolt	0.4734	30.3
TiRex	0.4735	30.3
Chronos-2	0.4799	29.3
statistical		
AR	0.4789	29.5
ARIMA(1,1,1)	0.5074	25.3
Theta	0.5104	24.9
Naive persistence	0.6792	0.0

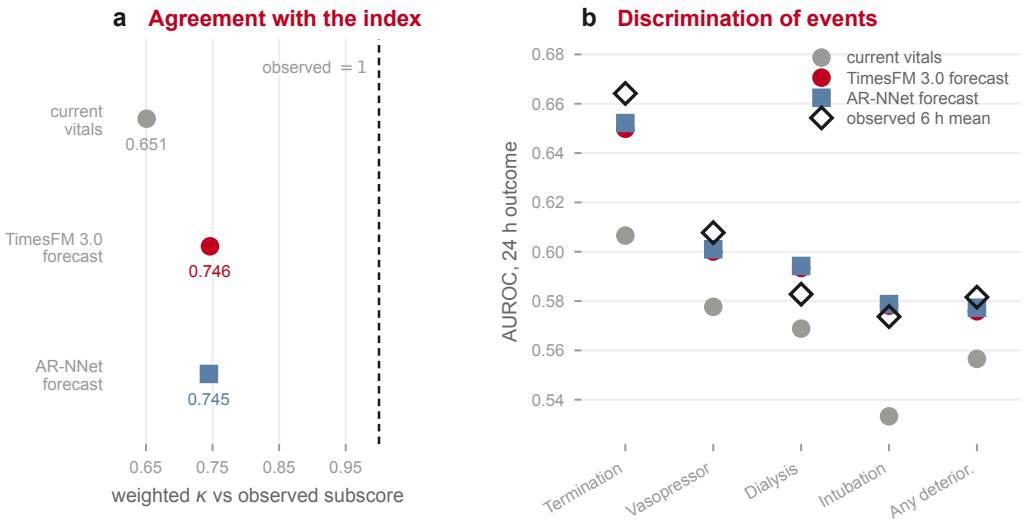


Figure 7: Two ground truths on one pipeline. **(a)** agreement with the risk subscore of the observed horizon mean (Stage 1), where the observed state is 1 by construction; **(b)** discrimination of recorded events in the next 24 h (Stage 2), where it is an ordinary comparator and is beaten on two of five outcomes.

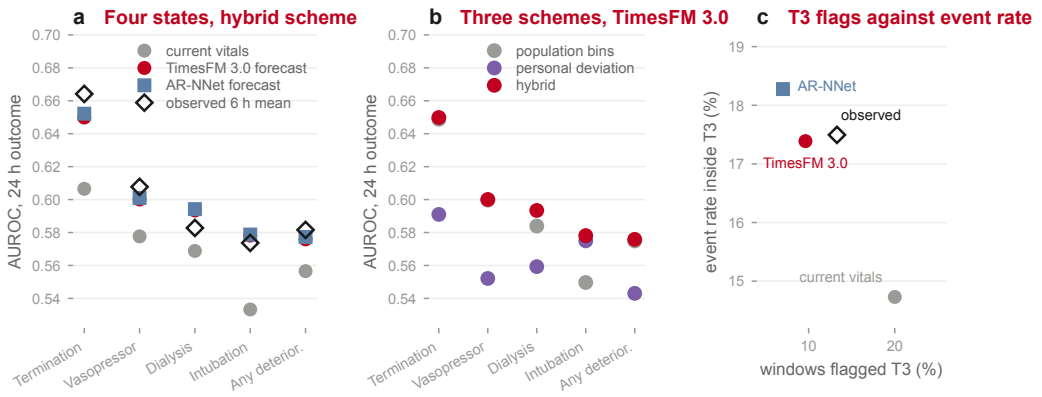


Figure 8: Triage against recorded outcomes in the following 24 h. **(a)** the four states under the *hybrid* scheme; **(b)** the three binning schemes on the TimesFM forecast, where personalisation wins on intubation alone, the one outcome preceded by within-patient change rather than absolute severity; **(c)** the alarm budget, T3 share against the event rate inside T3.

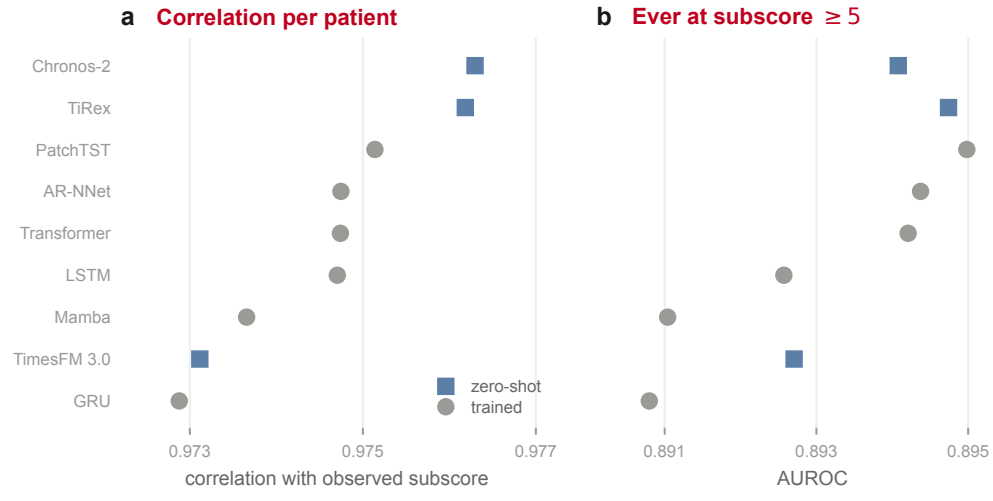


Figure 9: The risk read-out aggregated to the individual patient, nine models. **(a)** correlation between each patient's mean forecast subscore and the observed one; **(b)** AUROC for whether a patient is ever at subscore ≥ 5, where the whole spread across nine models is 0.0042. Models are in the same order in both panels: the ordering does not survive the change of metric.

Table 6: Triage against recorded outcomes in the following 24 h, AUROC, under the *hybrid* scheme. *current* is triage on the vitals observed at the forecast origin; the two forecast columns are triage on the 6 h forecast; *observed* applies the same rule to the observed 6 h mean and is a ceiling only for Risk Index calibration, not for these outcomes (Section 5). Each outcome carries its own eligibility mask, which is why n differs by row.

Outcome	n	**events**	**current**	**TimesFM**	**AR-NNet**	**observed**
Termination	542,987	8,266	0.607	0.650	0.652	0.664
Vasopressor start	436,483	30,270	0.578	0.600	0.601	0.608
Dialysis start	495,840	8,021	0.569	0.593	0.594	0.583
Intubation	259,492	22,933	0.533	0.578	0.579	0.574
Any deterioration	542,987	61,916	0.557	0.576	0.577	0.582

Table 7: Deteriorations recovered over triage on the currently observed vitals, at a forced 10% alarm budget on the window set common to all three horizons. Recall is the share of the $39{,}169$ events that fall inside the flagged 10%; Δ recall is in percentage points against the 12.57% reference. Intervals are percentile 95% from $1{,}000$ patient-level bootstrap resamples; * marks an interval excluding zero.

state	recall	Δ recall (pp)	relative gain	95% CI
current chart (reference)	12.57%	—	—	—
zero-shot TimesFM forecast				
H=6 h	13.34%	+0.78	+6.2%*	[+1.4, +10.7]
H=12 h	13.30%	+0.73	+5.8%*	[+1.1, +10.6]
H=24 h	13.04%	+0.47	+3.7%	[−1.4, +8.9]
trained AR-NNet forecast				
H=6 h	13.00%	+0.44	+3.5%	[−0.8, +8.0]
H=12 h	12.97%	+0.40	+3.2%	[−1.1, +7.9]
H=24 h	12.90%	+0.34	+2.7%	[−1.9, +8.1]
trained GRU forecast				
H=6 h	12.80%	+0.24	+1.9%	[−2.3, +6.5]
H=12 h	12.75%	+0.19	+1.5%	[−2.8, +6.5]
H=24 h	12.55%	−0.01	−0.1%	[−5.2, +5.4]
smoothing controls, no forecast				
mean of last 6 h charted	12.70%	+0.14	+1.1%	[−2.2, +4.5]
mean of last 24 h charted	12.44%	−0.12	−1.0%	[−6.5, +4.5]
observed-state comparator				
observed 6 h mean	13.74%	+1.18	+9.4%*	[+5.7, +13.1]

A.2 Patient-clustered intervals at a forced alarm budget

The budget-matched gains in the main text rank alarms by triage level at a budget of 9.6% of windows, TimesFM's own T3 share at H=6 h, applied unchanged to every state and horizon, so under level ranking that model alone needs no within-level tie-break. Their intervals, and those of the subscore-ranked variant, come from a patient-clustered paired bootstrap ($2{,}000$ resamples over the $4{,}403$ patients at H=6 h, seed 20260918). This appendix repeats the comparison under a second construction, which also supports paired contrasts across every model and horizon. The scoring code is the same, the identical four-parameter hybrid subscore, verified to reproduce the T3 shares and event rates of Section 6 before anything else is computed, and three things change. The budget is forced to exactly 10% of windows, with metrics taken as the expectation under uniform tie-breaking at the cut. Every state is scored on the $341{,}290$ windows common to H=6, 12 and 24 h ($39{,}169$ termination events over $3{,}046$ patients, $1{,}227$ of them with an event). Uncertainty comes from $1{,}000$ bootstrap resamples over patients, reported as percentile 95% intervals. The forced budget and the common window set both cost effect size, so the point estimates here are smaller than the level-ranked ones and the two must not be read as one number.

Table 7 is the result and three readings matter. The TimesFM forecast separates from the currently observed vitals at H=6 and H=12 h and not at H=24 h (Chronos-2 and TiRex still do; Table 4). The trained models do not separate at any horizon, which is the forced-budget counterpart of Section 8. The two smoothing controls, the same index scored on the mean of the last 6 or 24 h of observed values with no forecasting, both cover zero, so what the forecast contributes is the future state and not a quieter reading of the present. On this common set 131 patients carry a termination-bearing window.

A.3 Outcome connection to Risk Indexing

How the labels are built. For a window with a 48 h look-back ending at the forecast origin t_0, the lookahead interval is $[t_0,\ t_0+24\,\mathrm{h})$. It opens at t_0 and does not move with the forecast horizon. Support events for each modality (vasopressor, invasive ventilation, dialysis) are merged into episodes with a 12 h gap threshold and an episode *onset* counts only if no episode of that modality was active in the preceding gap. Each onset outcome is then restricted to windows where that modality was not



Figure 10: The patient of Figure 2: each scored vital over the 48 h look-back (solid line) and the observed 6 h horizon (dotted line), with its NEWS2 bands shaded behind the trace and the point value of each band on the right. The red segment is the forecast's horizon mean, which is the quantity the index bins. The figure shows why the index is coarse: respiratory rate spends the whole look-back inside one band, so none of that variation reaches the score, while systolic blood pressure crosses a boundary in the horizon and moves a point.

already running at t_0, since a therapy that is already in progress cannot newly start. This is stricter than flagging every administration record on purpose: repeated vasopressor starts within one stay would make the label read "this patient is currently on norepinephrine", a state recoverable from the concurrent blood-pressure trace, which would mess every AUROC result. Termination is a patient discharge inside the lookahead window. Termination event before the lookahead closes are flagged, so an unobserved negative is identifiable rather than silently counted.

What the termination label is and is not. "Termination" is used throughout as shorthand for this label and two properties of it need to be stated. First, it is a *hospital* discharge disposition, which puts it nearer hospital mortality than the in-ICU mortality that multi-centre ICU benchmarks score. YAIB takes a MIMIC death and then locates it to the unit through ward transfers, counting it only if the last ward was an ICU (van de Water et al., 2024). Second, MIMIC does not itself count a hospice discharge as a death, since such a patient is alive at discharge; our termination label does count it, so this label is broader than the database's own flag, on the view that a transition to hospice within the lookahead is an end-of-life event a deterioration score should not record as a negative. In-hospital mortality *or* hospice discharge is among the outcome measures under consideration for CDC and CMS sepsis benchmarking and it agrees with 30-day mortality substantially better than in-hospital mortality alone ($\kappa = 0.815$ against 0.698), precisely because a hospice discharge otherwise removes a patient from the count. The broadening is small and the result does not turn on it. Of the 8,266 complete-case terminations, 7,896 (95.5%) are Expired discharges and 370 (4.5%) Hospice. Rescoring the outcome on Expired discharges alone, at a forced 10% budget on the $542{,}987$ complete-case

windows at $H{=}6$ h (not the common window set of Section A.2), moves every mol's gain by at most 0.6 points (TimesFM $+18.8 \rightarrow +19.3\%$, AR-NNet $+23.4 \rightarrow +23.4\%$), with no interval changing sign and no change to the ordering in which the trained models lead on this outcome.

What the quantisation costs. The four NEWS2 rows used were checked row for row against the published NEWS2 chart (Royal College of Physicians, 2017), so the limitation of the four-parameter index lies in the three missing items and not in the four present ones. What the rows lose is the resolution of their bins. Binning is lossy by construction and the loss is measurable. Across the nine risk-scored models at $H{=}6$ h, the forecast subscore equals the observed subscore exactly in 38.8-40.4% of windows and falls within one point in 75.2-76.7%. For the currently observed vitals the same figures are 31.6% and 66.5%. Decomposing the disagreement, of the 59.6-61.2 percentage points of windows the forecast gets wrong, 36.2-36.4 points are a single-point crossing of a bin edge, about 60% of all disagreement, leaving 23.3-24.8 points, some 40%, at a distance of two or more. The current vitals are the state least concentrated on the boundary (51% of their disagreement is one point), which is consistent with their being wrong by more than a bin rather than marginally. The quadratic weighting in κ is what keeps the boundary crossings from dominating the statistic. A one-point error contributes a ninth of the penalty of a three-point error, which is why κ compresses the spread that squared error shows rather than amplifying it.

Two consequences follow and they point in opposite directions. The compression is partly a quantisation effect. With 76% of windows already within one point there is little room for a better forecaster to distinguish itself on this statistic, and the 1.6% κ spread over nine models should be read with that ceiling in mind rather than as evidence that the models are clinically interchangeable. The triage results do not inherit that ceiling, because they are scored against recorded events rather than against the index, and the event-anchored comparison is where the models separate (Section A.2). In short, the index read-out compresses and the outcome read-out discriminates, and only the second is evidence about clinical utility.

A.4 Why the two scores disagree: level and excursion

We decompose the observed trajectory into a horizon mean (the *level*) and deviations from it (the *excursion*), decompose the forecast the same way and regress each observed component on its predicted counterpart. This gives a Mincer-Zarnowitz coefficient per component that factorises accuracy over boldness. Every model recovers the level (0.982-1.009 trained, 0.855-0.867 zero-shot) and none predicts the excursion (correlations 0.030-0.095; Figure 11a). Squared error charges a model twice for unearned movement, once for the miss and once for the variance, while an ordinal band is largely blind to the second charge. This is why the shrinkage that squared error rewards is what a binned index punishes.

What makes the baseline strong: sensing density Assume-no-change is a hard baseline in the ICU because the ICU is densely monitored. Ageing its last observed value while leaving every model's look-back intact, its agreement falls from 0.651 at hourly observations to 0.460 at daily and the 6 h forecast's margin grows from $+14.6\%$ to $+62.4\%$ (Figure 11b). This is a deployment contrast and not a clean skill comparison, since only the baseline is degraded. Two properties of the read-out bound what it can see. Scored against a single spot hour instead of the horizon mean, every model loses 18-20% of its agreement, and 49.6% of windows contain an hour at subscore ≥ 5 while only 18.1% have a horizon mean that high, so averaging hides $2.7\times$ of what a bedside alarm sees. The forecast may earn its keep where sensing is sparse, which is the transfer these protocols are meant to enable.

A.5 Length of Stay

Section 6 scores the index against deterioration. Length of stay is a different kind of outcome, read by a bed manager rather than a bedside clinician and a predictor for it exists that reads no vital sign. This appendix asks whether the same index, unchanged, ranks patients on it.

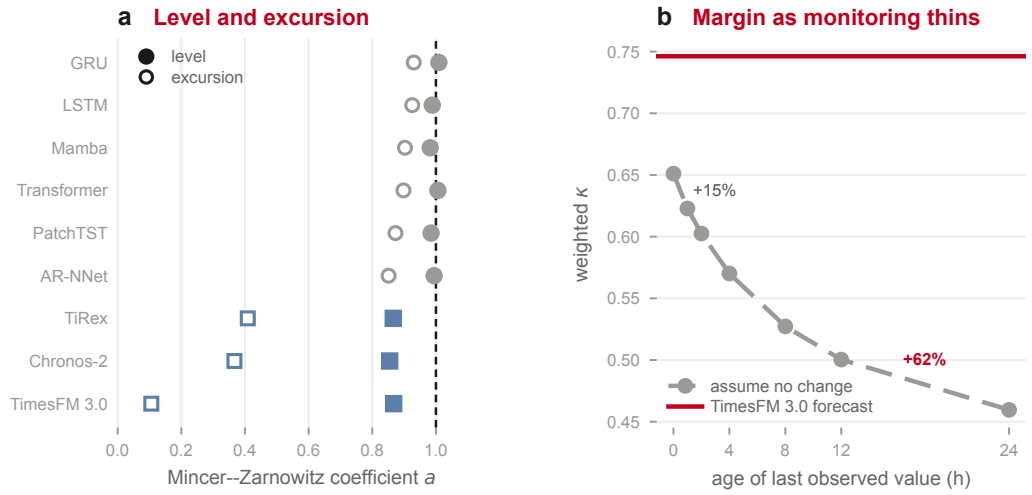


Figure 11: **(a)** level and excursion coefficients at 6 h in MIMIC; **(b)** agreement for assume-no-change as its last observed value is aged, against the 6 h forecast held fixed.

A.6 Outcome and control.

ICU episodes are built from the CLIF `adt` table; t_0 falls inside one for 83.4% of windows, 526,866 after the complete-case mask. Remaining hours cannot be the outcome directly, because a stay ends two ways that pull in opposite directions: among survivors a sicker patient stays longer, but a non-survivor patient leaves at once and in this cohort the pooled quantity runs the wrong way (median time remaining in the stay was 119 h for stays ending in termination or hospice and 91 h for live discharges). The outcome is therefore *still in the ICU at* $t_0 + 72\,h$, with termination before the mark negative, the question of three-days capacity suggested by the authors of YAIB's is the useful one (van de Water et al., 2024). Figure 12 adds a 24 h mark and a survivor-restricted version with the same reading. Scoring is the machinery of Section A.2 unchanged, and the published T3 shares are reproduced before it runs. The control is the clock: hours already elapsed in the ICU at t_0, quantile-binned to the same thirteen levels, so it enters the same triage machinery on the same footing.

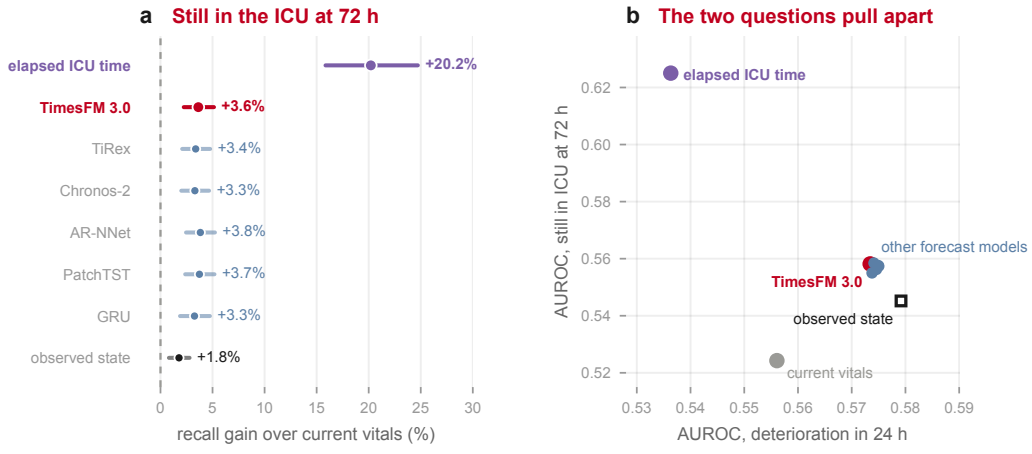


Figure 12: **(a)** recall gain over currently observed vitals at a 10% alarm budget on *still in the ICU at* 72 *h*, with subject-clustered intervals; **(b)** AUROC on deterioration against AUROC on remaining stay, one marker per state. The elapsed-time control sits in the opposite corner from every vitals-based state.

Table 8: **Length of stay, against the paper's own deterioration outcome.** $H=6$ h, hybrid bins, on the complete-case windows whose origin t_0 falls inside an ICU episode (n per column below). *Top:* recall at a forced 10% alarm budget, relative to the current chart, with a 95% interval from $1{,}000$ patient-level resamples; * marks intervals clear of zero. *Bottom:* AUROC, which fixes no budget. Every vitals-based state beats the chart on remaining stay by a few per cent, with no trained-versus-foundation ordering; the control, which reads no vital sign, beats it by twenty on both 72 h columns and falls below it on deterioration.

	resource outcome			**clinical outcome**
state	still in ICU at 24 h	still in ICU at 72 h	≥72 h more, survivors	deterioration in 24 h
recall at a 10% alarm budget, relative to current vitals (%)				
current chart	10.19%†	10.50%†	10.89%†	14.03%†
observed state	+0.6 [0.2, 1.1]*	+1.8 [0.8, 2.8]*	+5.0 [3.8, 6.3]*	+14.5 [11.7, 17.3]*
TimesFM 3.0	+1.5 [0.9, 1.9]*	+3.6 [2.2, 5.1]*	+6.1 [4.7, 7.8]*	+8.7 [5.3, 12.0]*
Chronos-2	+1.3 [0.9, 1.8]*	+3.3 [2.0, 4.7]*	+5.5 [4.1, 7.2]*	+9.6 [6.4, 12.6]*
TiRex	+1.4 [0.9, 1.8]*	+3.4 [2.1, 4.8]*	+5.6 [4.2, 7.1]*	+10.0 [6.8, 12.8]*
GRU	+1.4 [0.8, 1.8]*	+3.3 [1.9, 4.7]*	+6.1 [4.6, 7.8]*	+7.5 [4.2, 10.8]*
AR-NNet	+1.4 [1.0, 1.9]*	+3.8 [2.5, 5.2]*	+6.5 [5.1, 8.3]*	+8.3 [5.0, 11.4]*
PatchTST	+1.4 [0.9, 1.8]*	+3.7 [2.4, 5.2]*	+6.4 [4.9, 8.1]*	+8.0 [4.5, 11.2]*
elapsed ICU hours *(control)*	**+5.9 [4.8, 7.1]***	**+20.2 [15.9, 24.7]***	**+19.8 [14.6, 24.5]***	**-11.6 [-21.1, -1.7]***
AUROC				
current chart	0.529	0.524	0.531	0.556
observed state	0.555	0.545	0.560	0.579
TimesFM 3.0	0.565	0.558	0.572	0.573
Chronos-2	0.562	0.555	0.568	0.574
TiRex	0.564	0.556	0.569	0.575
GRU	0.567	0.558	0.574	0.574
AR-NNet	0.565	0.558	0.572	0.575
PatchTST	0.565	0.557	0.572	0.575
elapsed ICU hours *(control)*	0.622	0.625	0.644	0.536
n	526,866	526,866	390,306	526,866
prevalence	0.868	0.618	0.596	0.112

†absolute recall of the current-chart row, which the relative figures above are taken against.

Two further readings are worth recording. The trained-versus-foundation ordering of Section 8 does not appear in these point estimates: the six forecast models span $+3.3$ to $+3.8\%$ on the 72 h outcome with a trained network nominally on top, where on deterioration in the same table all three zero-shot models lead all three trained ones. Paired model-versus-model contrasts were not run on the stay outcomes, so this is an observation about the ordering and not a test of it. It is consistent with the family split being about deterioration-relevant forecast content rather than a property of the scoring machinery, on one outcome. The observed state also changes role again, as it does between the two stages of Section 5: it leads the forecast models on deterioration ($+14.5\%$) and trails all of them on remaining stay ($+1.8\%$ against $+3.3$ to $+3.8$). Knowing the next six hours of vitals is not a better answer to the capacity question than forecasting them, and neither is a good one.