

Dynamic prediction of early hepatic deterioration in critically ill patients with cirrhosis: model development in MIMIC-IV and transportability assessment across four international intensive care databases

Zhihui Zhao

Heping Hospital Affiliated to Changzhi Medical College

Zhenhua Zhu

Heping Hospital Affiliated to Changzhi Medical College

Hongli Gao

gaohongli@sxbqeh.com.cn

Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Third Hospital of Shanxi Medical University

Research Article

Keywords: cirrhosis, intensive care unit, dynamic prediction, landmark analysis, machine learning, external validation, model transportability, bilirubin, hepatic deterioration, MIMIC-IV

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10929847/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Dynamic prediction of early hepatic deterioration in critically ill patients with cirrhosis: model development in MIMIC-IV and transportability assessment across four international intensive care databases

Zhihui Zhao¹, Zhenhua Zhu², Hongli Gao^{3*}

1 Department of Pharmacy, Heping Hospital Affiliated to Changzhi Medical College, Changzhi, Shanxi, China

2 Department of Intensive Care Unit, Heping Hospital Affiliated to Changzhi Medical College, Changzhi, Shanxi, China

3 Department of Pharmacy, Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Third Hospital of Shanxi Medical University, Tongji Shanxi Hospital, Taiyuan, Shanxi, China

* Corresponding author: Hongli Gao, gaohongli@sx bqeh.com.cn

Abstract

Background: Patients with cirrhosis admitted to the intensive care unit (ICU) are vulnerable to early biochemical deterioration of liver tests, yet existing prediction models are single-centre and lack external assessment. We rebuilt a multi-landmark dynamic prediction framework on MIMIC-IV 3.1 and evaluated its transportability across four independent ICU databases.

Methods: This retrospective study developed models in 6,703 MIMIC-IV ICU stays with ICD-coded cirrhosis and evaluated a portable model in 63,268 external stays (eICU-CRD, AmsterdamUMCdb, SICdb, NWICU). The primary endpoint was composite laboratory-defined hepatic deterioration (new elevation of ALT, AST, total bilirubin, or ALP above fixed thresholds after hour 24 among stays below threshold at baseline); a portable total-bilirubin endpoint was defined identically in all databases, and a universal worsening endpoint also captured progression among patients already above threshold. Landmark models (6-48 h landmarks; 24-72 h horizons) were trained with logistic regression and LightGBM under patient-grouped five-fold cross-validation, benchmarked against naive threshold-proximity rules, and applied externally without refitting.

Results: Composite deterioration occurred in 971 of 5,509 gradable stays (17.6%). At the primary 24 h landmark / 48 h horizon, the model achieved an out-of-fold AUROC of 0.817 (95% CI 0.800-0.834) versus 0.699 for the strongest naive rule (increment +0.118, 95% CI 0.095-0.142). The portable model (development AUROC 0.886) transported without

refitting to eICU (0.897), SICdb (0.880), AmsterdamUMCdb (0.876), and NWICU (0.901); raw risks exceeded observed rates four- to six-fold and were restored by a single intercept correction. Within diagnosis-matched eICU cirrhosis patients the portable model retained AUROC 0.808-0.848, and a multivariable cirrhosis-probability phenotype tracked coded diagnoses externally (AUROC 0.888-0.965). The universal worsening endpoint reached AUROC 0.895 in development, but its label was partly measurement-frequency dependent and its construct did not transport to eICU cirrhosis patients; it is presented as hypothesis-generating.

Conclusions: Early laboratory dynamics predicted impending hepatic deterioration with discrimination that clearly exceeded naive rules, and the portable model discriminated consistently across four health systems while requiring local intercept recalibration. These findings support prospective evaluation of laboratory-trajectory screening in ICU patients with liver vulnerability.

Keywords: cirrhosis; intensive care unit; dynamic prediction; landmark analysis; machine learning; external validation; model transportability; bilirubin; hepatic deterioration; MIMIC-IV

Introduction

Patients with cirrhosis admitted to the intensive care unit (ICU) operate near the limits of their hepatic reserve. Superimposed insults – infection, shock, drug exposure, and procedural stress – frequently precipitate biochemical deterioration of liver tests that precedes overt organ failure and is associated with substantial mortality [1-5]. Because transaminase and bilirubin trajectories evolve over hours, a clinically useful early-warning tool must be dynamic: it should repeatedly re-estimate risk as data accumulate, rather than rely on a single static assessment at admission.

The landmark dynamic prediction framework addresses this requirement by conditioning on survival and information availability at successive time points (landmarks) and predicting events within forward time windows (horizons) [6,7]. The approach has been applied to mortality and acute kidney injury prediction in critical illness [8,9]. We previously explored its application to a laboratory-defined, DILI-like hepatic deterioration endpoint in a small HBV-cirrhosis cohort (unpublished data), where discrimination was moderate, confidence intervals were wide, monitoring-intensity variables dominated feature importance, and

prescription-based drug-exposure features contributed no incremental information. That study – like most EHR-based hepatic injury prediction work – was limited by a single centre, a small event count, and the absence of any external assessment [10,11].

Three barriers have hindered progress in this field. First, adjudicated drug-induced liver injury (DILI) is rare and cannot be ascertained at scale in retrospective ICU data, forcing reliance on laboratory surrogate endpoints whose purity is contested [12,13]; benchmark resources for DILI risk remain preclinical or drug-level rather than patient-level [11]. Second, single-database studies cannot distinguish genuine signal from local practice patterns, including laboratory monitoring behaviour. Third, transportability across health systems – the decisive question for any deployable early-warning system – has rarely been examined for hepatic endpoints.

We therefore rebuilt the prediction framework using the updated MIMIC-IV version 3.1 database as a development cohort including patients with cirrhosis of all aetiologies, and four independent ICU databases (eICU-CRD, AmsterdamUMCdb, SICdb, and NWICU) as external populations spanning the United States, the Netherlands, and Austria. Our objectives were: (1) to quantify the performance of multi-landmark models for a composite hepatic deterioration endpoint in a large cirrhosis ICU cohort, benchmarked against naive threshold-proximity rules to exclude task circularity; (2) to develop a parsimonious portable model restricted to laboratory variables available in every database and test its unmodified transportability in 63,268 external ICU stays, overall and within a laboratory-defined liver-vulnerable phenotype; (3) to quantify outcome-label heterogeneity through a Sepsis-3-based sensitivity analysis; and (4) to anchor the surrogate endpoint by relating predicted risk to mortality in every database. Because the external databases lack discharge diagnosis tables, this design constitutes a transportability assessment in general ICU populations rather than an aetiology-matched external validation, and we report it as such.

Methods

Data sources

This retrospective, observational study used five publicly available, de-identified critical care databases: MIMIC-IV version 3.1 (Beth Israel Deaconess Medical Center, Boston, USA; 2008-2022) [14], eICU Collaborative Research Database version 2.0 (multi-centre, USA) [15], AmsterdamUMCdb version 1.0.2 (Amsterdam UMC, the Netherlands) [16], the

Salzburg Intensive Care database (SICdb) version 1.0.8 (University Hospital Salzburg, Austria) [17], and the Northwestern ICU database (NWICU) version 0.1.0 (Northwestern Memorial HealthCare, USA) [18]. All databases were accessed through a harmonised concept-level extraction layer that maps source tables to standardised variables with explicit unit normalisation (e.g., total bilirubin to mg/dL) and time anchoring relative to ICU admission; key conventions of this layer, including unit conversion rules, outlier flagging, and the assignment of hospitalisation-level laboratory events to ICU stays by time-window overlap, are documented in Additional file 1. Each source database was released under a data use agreement permitting credentialed research use; the institutional review boards of the source institutions approved the releases and waived informed consent for de-identified data. This study is reported according to the TRIPOD+AI statement [19,20].

Development cohort

The development cohort comprised all ICU stays in MIMIC-IV 3.1 whose hospitalisation carried an ICD-9 or ICD-10 discharge diagnosis of cirrhosis or hepatic fibrosis (ICD-10 K70.2-K70.4, K74.x; ICD-9 571.2, 571.5, 571.6); the K74.x block includes hepatic fibrosis without explicit cirrhosis (K74.0-K74.2) and primary biliary cholangitis or primary sclerosing cholangitis (K74.3-K74.5), which we retain for comparability with our previous study. Etiology flags were derived for hepatitis B (B16.x, B18.0, B18.1, B18.8, B18.9, 070.2x, 070.3x), hepatitis C (B17.1, B18.2, 070.4x, 070.5x), and alcohol-related liver disease (K70.x, 571.0-571.3). Where a hospitalisation contained several ICU stays, each stay was analysed separately with laboratory events assigned to the ICU stay whose time window contained them.

Outcome definitions

Two laboratory-based dynamic endpoints were prespecified. Endpoint A (development; full liver panel) replicated the definitions of our previous single-centre analysis (unpublished data): for each of ALT, AST, total bilirubin (TBIL), and ALP, the baseline value was the earliest measurement within 24 h of ICU admission; a soft deterioration event in a given analyte was the first value at or above its threshold (ALT ≥ 120 U/L, AST ≥ 120 U/L, TBIL ≥ 2.4 mg/dL, ALP ≥ 240 U/L) occurring after hour 24 among stays whose baseline was below the threshold; a strict event additionally required at least two abnormal measurements with a peak at least doubling the baseline. The composite endpoint was any soft event across the four analytes. Endpoint B (portable) applied the same logic to TBIL

alone, the only liver-specific analyte harmonised across all five databases. Stays were considered gradable for an endpoint if at least one constituent analyte had a baseline value below its threshold. Two design consequences follow and are addressed explicitly: conditioning on a below-threshold baseline excludes from each analyte's event definition those stays already abnormal at baseline (56.4% for TBIL among gradable stays), so the endpoints capture new deterioration rather than further worsening of established abnormalities; and because events can occur only after hour 24, short landmark-horizon combinations have attenuated at-risk time.

Universal worsening endpoint (extension for already-abnormal baselines)

Because the conditional endpoints exclude patients already above threshold at baseline – 56.4% of gradable stays for bilirubin – we additionally defined a universal worsening endpoint family that covers the full spectrum of baseline abnormality. For each analyte, a universal event was the union of: (i) new crossing, as above (baseline below threshold, value at or above threshold after hour 24); and (ii) progression, applicable when the baseline was at or above threshold: any post-24 h value at least twice the baseline with an absolute increment of at least 40 U/L (ALT, AST), 1.0 mg/dL (TBIL), or 80 U/L (ALP), guarding against trivial doubling of small values. The universal composite (endpoint A2) and universal bilirubin endpoint (endpoint B2) were labelled and modelled identically to the conditional endpoints, with gradability requiring any baseline measurement in the first 24 h rather than a below-threshold baseline. The event time was the first hour of the first qualifying value, and risk-set construction was unchanged.

Cirrhosis-probability phenotype

To approximate aetiology-matched external validation despite the absence of diagnosis tables outside MIMIC-IV, we trained a multivariable cirrhosis-probability phenotype on 45,621 MIMIC-IV ICU stays (6,620 with ICD-coded cirrhosis; 40,000 randomly sampled non-cirrhosis stays) using only first-24 h portable laboratory summaries (minimum, maximum, mean, last value of the nine portable analytes) plus age and sex, with patient-grouped five-fold cross-validation. A restricted variant excluding all liver-related analytes (bilirubin, INR, prothrombin time, aPTT) – retaining creatinine, platelet count, lactate, white cell count, haematocrit, age, and sex – quantified how much of the phenotype's power depends on liver tests and therefore its degree of circularity with liver-focused endpoints. Both variants were applied to the external databases, and portable-model performance was

reported within the top 30% of predicted cirrhosis probability versus the remainder. A simpler fixed-threshold phenotype – first-24 h minimum platelet count below $150 \times 10^9/L$ – is retained alongside the probability phenotype for continuity with our previous report and for bedside transparency; the probability phenotype is the validated upgrade used for stratified reporting.

True diagnosis-matched validation using raw diagnosis tables

For three external databases, the raw source diagnosis tables were available and were analysed directly: the eICU diagnosis table (2,710,672 rows; hierarchical diagnosis strings plus ICD-9 codes; 2,512 cirrhosis-flagged unit stays), the SICdb case table (ICD-10 main admission diagnosis and diagnosis text; 73 flagged cases, reflecting the fact that only primary admission diagnoses are recorded), and the NWICU ICD diagnosis table (201,355 rows mapped from hospitalisation to ICU-stay identifiers; 502 flagged stays). AmsterdamUMCdb's diagnosis table was not available locally and that database therefore remains validated through the phenotype approach only. Cirrhosis flags used ICD-9 571.2/571.5/571.6, ICD-10 K70.2-K70.4 and K74.x, and text keywords (cirrhosis, liver fibrosis, oesophageal varices, portal hypertension, hepatic failure). The same flags served two purposes: restricting the external validation populations to diagnosis-matched cirrhosis patients, and externally validating the cirrhosis-probability phenotype itself against coded diagnoses. Because the keyword list included 'hepatic failure' – which captures acute liver failure without cirrhosis – all diagnosis-matched eICU results are reported for both the full and a strict keyword variant excluding that term. Administrative censoring was quantified as the proportion of at-risk stays discharged from the ICU before completing the 72 h outcome window, and the primary analysis was repeated restricted to window-completers. Calibration within the diagnosis-matched cirrhosis subgroup (mean predicted risk, intercept-only correction, Brier score) was reported separately from the all-comer correction.

Landmark framework and risk sets

Prediction landmarks were set at 6, 12, 24, and 48 h after ICU admission with horizons of 24, 48, and 72 h (12 combinations). At each landmark, the risk set comprised stays still in the ICU at the landmark, gradable for the endpoint, and free of the endpoint event before the landmark; the label indicated whether the first event occurred within (landmark, landmark + horizon]. Stays discharged from the ICU before completing a horizon were treated as event-free (right-censoring at ICU discharge).

Features

The development feature set (218 variables) summarised data available up to each landmark: for each of 18 routine laboratory analytes (ALT, AST, ALP, TBIL, direct bilirubin, INR, albumin, creatinine, platelet count, lactate, white cell count, haematocrit, sodium, potassium, bicarbonate, urea nitrogen, prothrombin time, aPTT) ten summary variables (count, first, last, minimum, maximum, mean, change, slope, first and last measurement time); static variables (age, sex); severity scores (MELD [21], first-day SOFA and its liver component, SAPS-II, OASIS, Charlson index); counts of exposure to categories of high-risk hepatotoxic drugs administered before the landmark (acetaminophen, antituberculosis agents, amiodarone, antiepileptics, azole antifungals, macrolides, statins, linezolid), identified from ICU administration records only – oral outpatient agents such as amoxicillin-clavulanate, methotrexate, and valproate are not captured; and a laboratory missingness count. The portable feature set restricted the same construction to nine analytes available in all databases (TBIL, INR, creatinine, platelet count, lactate, white cell count, haematocrit, prothrombin time, aPTT) plus age and sex.

Models, benchmarks, and validation strategy

Logistic regression (median imputation, standardisation, class weighting) and LightGBM (150 trees, learning rate 0.03, seven leaves, depth 3, subsampling 0.8, L1 0.5, L2 1.0, class-weighted [22]; hyperparameters fixed a priori from our previous study, with no tuning on the present data) were compared at every landmark-horizon combination using five-fold GroupKFold cross-validation grouped by patient, evaluated out-of-fold by AUROC, AUPRC, and Brier score. The 24 h landmark / 48 h horizon combination was designated primary before analysis, carried over from the design of our previous study on the basis of clinical actionability; all 12 combinations are additionally reported in full. Bootstrap percentile confidence intervals (1,000 resamples) are reported for the primary estimate, all subgroup and external AUROCs, and for incremental comparisons.

Because the portable endpoint is defined by a bilirubin threshold, model performance could in part reflect threshold proximity rather than trajectory learning. We therefore benchmarked every model against naive rules computable at the bedside without machine learning: for the composite endpoint, the maximum across the four analytes of the last-value-to-threshold ratio (proximity) and of the change-to-threshold ratio (rise); for the portable endpoint, the last TBIL value, the TBIL change, and a fixed rule flagging last TBIL ≥ 1.8 mg/dL or a

rise ≥ 0.3 mg/dL. Incremental performance was quantified as the paired difference in AUROC with bootstrap confidence intervals. For the universal endpoints, a mechanistically motivated trajectory-extrapolation rule was additionally benchmarked: the first-24 h slope of each analyte was linearly extrapolated to hour 72, scored against the applicable event threshold (the fixed threshold for below-threshold baselines, twice the baseline for already-abnormal baselines), and the maximum score across analytes used as the naive risk score. Because progression-type events require at least one further measurement after hour 24 to be detectable, we additionally quantified the measurement-frequency confound of the universal endpoints: the AUROC of the raw post-24 h bilirubin draw count alone as a predictor of the universal event, and ablations of the universal models with all monitoring-intensity features removed.

Calibration was assessed by the logistic recalibration slope and intercept, by decile calibration plots [23], and – for the external databases – by an intercept-only correction (a single additive shift on the logit scale, the minimal adjustment a deploying site could apply) with corrected mean predicted risk and corrected Brier score. Decision curve analysis [24] was performed on Platt-recalibrated probabilities because class-weighted raw probabilities are overdispersed by construction; the raw curve is shown for comparison. Feature importance used SHAP values (TreeExplainer) [25]. Prespecified ablations compared feature subsets (static, scores, laboratory dynamics, laboratory dynamics without monitoring-intensity variables, medications, and combinations). Subgroup performance was assessed by cirrhosis aetiology (HBV, HCV, alcohol) with bootstrap confidence intervals.

The portable model (LightGBM on endpoint B, landmark 24 h, horizon 48 h) was trained on the MIMIC-IV cirrhosis cohort and applied without any refitting or threshold adjustment to the four external databases. Because the external databases lack discharge diagnosis tables in the harmonised layer, external cohorts comprised all ICU stays gradable for endpoint B; results are reported both overall and within a liver-vulnerable phenotype defined a priori as a minimum first-24 h platelet count below $150 \times 10^9/L$. This phenotype was first validated internally in MIMIC-IV against the ICD-coded cirrhosis cohort. Because the phenotype identifies acuity as well as chronic liver disease (thrombocytopenia from sepsis or haemodilution), subgroup event enrichment blends hepatic vulnerability with illness severity, and results are interpreted accordingly. To quantify outcome-label heterogeneity, we measured the proportion of endpoint-B events with a Sepsis-3 onset [26] within the preceding 48 h in each database, and repeated the primary analyses after excluding sepsis-

proximate events. To relate the surrogate endpoint to outcome, hospital mortality (30-day mortality in AmsterdamUMCdb, where hospital-mortality flags were unavailable) was tabulated by model-predicted risk quartile in each external database; these associations cannot separate endpoint validity from the severity-marker properties of the constituent laboratory variables. Analyses used Python (pandas, scikit-learn, LightGBM, SHAP).

Results

Cohort characteristics

Of 94,458 ICU stays in MIMIC-IV 3.1, 6,703 stays (5,584 hospitalisations; 4,182 patients) met the cirrhosis definition and formed the development cohort. Median age was 60 years (IQR 52-67), 64.4% were male, and aetiology flags indicated alcohol-related disease in 53.8%, hepatitis C in 12.4%, and hepatitis B in 3.0%. Median MELD was 23 (15-31), median first-day SOFA 7 (5-10), and hospital mortality was 24.5% (Table 1). Baseline liver-test abnormalities were common: among stays with a gradable baseline, total bilirubin was already at or above 2.4 mg/dL in 56.4%, AST at or above 120 U/L in 29.4%, and ALT in 15.7%. Among the 5,509 stays gradable for endpoint A, the composite endpoint occurred in 971 (17.6%), including 567 (10.3%) meeting strict definitions; the most frequent components were AST (443), ALT (345), INR (367, characterisation only), TBIL (328), and ALP (272) events.

Development performance and benchmark against naive rules

LightGBM outperformed logistic regression in all 12 landmark-horizon combinations (Table 2, Figure 1). Discrimination peaked at the 24 h landmark: AUROC 0.845 (AUPRC 0.466) with a 24 h horizon and 0.817 (AUPRC 0.478) with a 48 h horizon, against an event rate of 13.4% (613 events among 4,585 at risk). At the prespecified primary combination the bootstrap 95% CI for AUROC was 0.800-0.834 and for AUPRC 0.437-0.521. Performance declined at the 48 h landmark (AUROC 0.651-0.712), reflecting the heavily depleted and selected risk set of long-staying patients in whom the event has not yet occurred.

Naive benchmarking distinguished genuine trajectory learning from threshold proximity (Table 8). For the composite endpoint, the strongest naive rule was the gap-completion score (the fraction of the baseline-to-threshold distance covered within 24 h, maximised across the four analytes), achieving AUROC 0.699 – above the simpler proximity (0.585) and relative-rise (0.661) rules but still far below the model's 0.817; the incremental AUROC against the

gap-completion rule was +0.118 (bootstrap 95% CI 0.095-0.142, $p < 0.001$), and no fixed-threshold rule approached model performance. For the portable bilirubin endpoint the picture was more nuanced: the last TBIL value alone reached AUROC 0.864 and the fixed rule (last ≥ 1.8 mg/dL or rise ≥ 0.3 mg/dL) operated at sensitivity 0.742 / specificity 0.805 / PPV 0.343, while the model achieved 0.886 – an incremental AUROC of +0.032 (95% CI 0.014-0.052, $p = 0.002$). We therefore interpret the portable model as a modest refinement of an already informative bilirubin trajectory, whereas the composite endpoint requires genuine multi-analyte integration that naive rules cannot supply.

Primary model deep-dive (24 h landmark, 48 h horizon)

Predicted-risk quartiles were strongly separated (event rates 2.6%, 4.4%, 11.2%, and 35.3%; Figure 2), a 13.5-fold gradient between the top and bottom quartiles. Raw probabilities were overdispersed relative to observed rates (mean predicted 0.367 vs observed 0.134; recalibration slope 1.38, intercept -1.55), a known consequence of class weighting; Platt recalibration restored correspondence (Figure 2C). Decision curve analysis on recalibrated probabilities showed positive net benefit across threshold probabilities of 0.02-0.45, exceeding treat-all at thresholds above 0.15 (Figure 2D). SHAP analysis attributed the largest contributions to the dynamics of total bilirubin, AST, INR, and creatinine, together with their measurement counts (Figure 4).

In ablation analysis, laboratory dynamics alone reproduced the full model (AUROC 0.816 vs 0.817); static variables (0.508), severity scores (0.667), and hepatotoxic medication exposure (0.511) added no incremental discrimination (Table 4). Removing monitoring-intensity variables (measurement counts and timing) reduced AUROC only from 0.817 to 0.813, indicating that the predictive signal resides predominantly in the laboratory values themselves rather than in who is tested – a direct contrast with our previous small HBV-cirrhosis cohort, where the same removal degraded AUROC from 0.727 to 0.655; we note, however, that the very availability of serial trend variables remains weakly correlated with monitoring intensity. Performance was consistent across aetiologies: AUROC 0.813 (95% CI 0.698-0.914; 17 events) in HBV-related, 0.848 (0.801-0.894) in HCV-related, and 0.825 (0.797-0.851) in alcohol-related cirrhosis (Table 5).

Portable model and transportability across four international databases

The portable nine-analyte model for bilirubin deterioration achieved an out-of-fold AUROC of 0.886 (AUPRC 0.564) at the 24 h landmark / 48 h horizon in the MIMIC-IV cirrhosis

cohort (12.1% event rate). Applied without refitting to the four external databases (Table 3, Figure 3), it reached an AUROC of 0.897 (95% CI 0.887-0.907) in eICU (n=50,496; 970 events; 1.9%), 0.880 (0.847-0.908) in SICdb (n=5,455; 3.1%), 0.876 (0.841-0.906) in AmsterdamUMCdb (n=6,665; 2.0%), and 0.901 (0.802-0.976; 12 events, wide interval) in NWICU (n=652; sparse laboratory coverage). Discrimination was thus preserved across case mix, assay practice, and markedly different bilirubin sampling density (median sampling frequency differs several-fold between systems).

Calibration did not transport. Raw mean predicted risks (0.084-0.128) exceeded observed event rates (0.018-0.031) by roughly four- to six-fold, reflecting both class weighting and case-mix dilution in all-comer populations. An intercept-only correction – a single logit-scale shift of -1.79 to -2.39, estimable by any deploying site from local outcome data – restored calibration-in-the-large exactly and improved Brier scores in every database (e.g., eICU 0.026 to 0.017; AmsterdamUMCdb 0.043 to 0.017). Logistic calibration slopes after the intercept correction were 1.01 (SICdb) and 1.06 (AmsterdamUMCdb) – near unity – and 1.23 (eICU) and 1.32 (NWICU) – mildly under-confident (Table 3). We therefore recommend local intercept recalibration before any deployment, and caution that uncorrected raw outputs would over-trigger substantially.

The platelet-based liver-vulnerability phenotype was first validated internally: among 44,358 MIMIC-IV stays sampled for phenotype assessment, a first-24 h minimum platelet count below $150 \times 10^9/L$ identified ICD-coded cirrhosis with AUROC 0.798 (sensitivity 0.773, specificity 0.685, PPV 0.296 at approximately 15% prevalence). Applied externally, the phenotype enriched endpoint-B risk 2.5- to 4.3-fold: event rates in liver-vulnerable versus non-vulnerable subgroups were 4.0% versus 1.0% (eICU), 6.7% versus 1.6% (SICdb), and 4.2% versus 0.8% (AmsterdamUMCdb). Discrimination within subgroups was maintained (SICdb 0.884 vs 0.821) but not uniformly higher than in non-vulnerable patients (eICU 0.870 vs 0.884), consistent with the phenotype capturing severity as well as liver disease (Table 3).

Universal worsening endpoint: covering patients already abnormal at baseline

Under the universal definitions, 2,827 of 5,509 stays with any baseline (51.3%) experienced clinically significant worsening, of which 2,503 arose through progression in patients already at or above their analyte threshold and 971 through new crossing – confirming that the majority of the worsening burden in this cohort lies in the population the conditional

endpoints exclude. The universal bilirubin endpoint flagged 2,392 of 5,477 stays. At the 24 h landmark / 48 h horizon, the universal composite model (A2) achieved AUROC 0.895 (95% CI 0.884-0.905; AUPRC 0.908) among 4,671 at-risk stays with 2,674 events (57.2% event rate), and the universal portable model (B2) achieved 0.914 (0.906-0.922; AUPRC 0.891; 48.7% event rate); the high event rates reflect that clinically significant laboratory worsening within 48 h is the norm rather than the exception among cirrhosis ICU patients still at risk at 24 h, which shifts the model's role from rare-event detection to trajectory triage. Naive benchmarking assumed greater importance for this endpoint family: threshold proximity alone reached 0.862 (composite) and 0.892 (last bilirubin value), with model increments of +0.033 (0.026-0.041) and +0.050 (0.042-0.059) respectively – significant but modest, because level-proximity is intrinsically informative once already-abnormal patients dominate the risk set (Table 10). Transportability of the universal portable model was preserved, with external AUROCs of 0.844 (0.835-0.852) in eICU (n=55,351; 2.0% event rate), 0.825 (0.796-0.851) in SICdb (n=6,067; 3.2%), 0.853 (0.824-0.880) in AmsterdamUMCdb (n=7,072; 2.1%), and 0.814 (0.675-0.911) in NWICU (n=719; 12 events). Event composition, however, inverted across systems: in development, 86% of universal bilirubin events were progression-type (2,064 of 2,392), whereas detectable external events were 80-86% crossing-type (eICU 1,842 crossing vs 453 progression; SICdb 374 vs 61; AmsterdamUMCdb 311 vs 58) – the measured basis of the construct non-transportability demonstrated below.

Cirrhosis-probability phenotype and phenotype-restricted external validation

The multivariable cirrhosis-probability phenotype discriminated ICD-coded cirrhosis in MIMIC-IV with an internal AUROC of 0.910 (0.906-0.915); even the restricted variant excluding every liver-related analyte retained 0.868 (0.862-0.872), indicating a substantial non-liver laboratory signature of chronic liver disease (thrombocytopenia, severity-pattern chemistry, demographics) that is available for phenotyping even when liver tests are not. Applied to the external databases, the top 30% of the full phenotype's predicted cirrhosis probability concentrated universal-endpoint events 4- to 13-fold relative to the remainder (event rates 5.2% vs 0.6% in eICU, 7.9% vs 1.2% in SICdb, 5.7% vs 0.6% in AmsterdamUMCdb; Table 11), and the restricted variant enriched 2.9- to 4.4-fold while avoiding any dependence on liver tests. Within these high-probability strata – the closest available approximation to the target population – portable-model discrimination was attenuated relative to cross-population values (AUROC 0.646-0.748 across databases for the

full phenotype), as expected when case-mix breadth is removed, but remained above 0.64 in every database and above 0.74 for the restricted variant's strata. These analyses convert population mismatch from an unquantified limitation into a measured gradient, while diagnosis-matched validation in databases with coded diagnoses remains the definitive standard.

True diagnosis-matched validation

Analysing the raw diagnosis tables closed the loop on population validity (Table 12). In eICU, 591 diagnosis-matched cirrhosis stays entered the conditional-endpoint risk set and experienced 60 events – an event rate of 10.2%, closely matching the 12.1% of the MIMIC-IV development cohort – and the portable model retained good discrimination within these patients (AUROC 0.808, 95% CI 0.748-0.865), modestly below the non-cirrhosis value of 0.896 as expected under case-mix narrowing. SICdb and NWICU cirrhosis subgroups were too small for estimation (7 and 6 conditional-endpoint stays with 1-2 events each; 30 and 22 universal-endpoint stays), because SICdb records only primary admission diagnoses and NWICU laboratory coverage is sparse. The cirrhosis-probability phenotype itself was validated against coded diagnoses in all three databases: AUROC 0.888 (0.880-0.897) in eICU, 0.965 (0.946-0.979) in SICdb, and 0.961 (0.935-0.982) in NWICU, with the restricted liver-free variant reaching 0.816-0.912 – confirming that the phenotype stratification of Table 11 tracks true coded cirrhosis rather than an artefact of liver-test circularity.

The diagnosis-matched analysis also exposed the boundary of the universal endpoint. Within eICU cirrhosis patients, the universal model's AUROC fell to 0.414 (0.369-0.461) while remaining 0.854 in non-cirrhosis patients. The mechanism is informative: eICU measures total bilirubin sparsely (median approximately two draws per stay), so doubling-based progression in already-abnormal patients is largely undetectable, and the detectable events concentrate in low-baseline crossers – precisely the patients the development-trained universal model ranks lower. The universal endpoint family therefore depends on measurement density and should be deployed only where laboratory monitoring is at least daily; the conditional endpoint, whose label logic requires a single threshold-crossing value, is robust to sparse sampling.

Robustness analyses: monitoring confound, extrapolation benchmark, keyword sensitivity, and censoring

Three findings sharpened the interpretation of the endpoint families (Table 13). First, the universal endpoints carry a measurement-frequency confound: the raw count of post-24 h bilirubin draws alone predicted the universal bilirubin event at AUROC 0.807 (0.794-0.819), and the count of any liver-panel draws predicted the universal composite at 0.867 – confirming that detectability of doubling-based progression is partly a function of testing intensity. Removing all monitoring-intensity features from the universal models cost little discrimination (portable universal 0.914 to 0.906; composite universal 0.895 to 0.889), indicating that the confound operates through the event label rather than through the features: the models rank trajectories well, but the label itself is partially a measurement-frequency construct. Second, the trajectory-extrapolation naive rule – the strongest benchmark a bedside rule could plausibly achieve for the universal endpoint – reached only 0.574 (composite) and 0.607 (portable), far below the models, so the naive-benchmark margin for the universal endpoints remains +0.31 or more against every rule tested. Third, restricting the eICU cirrhosis flag to the strict keyword variant (excluding 'hepatic failure') left 471 stays and 43 events in the conditional risk set with AUROC 0.848 (0.790-0.897) – stronger than the inclusive definition – while the universal endpoint remained inverted (0.439, 0.393-0.490). Within eICU cirrhosis patients having three or more post-24 h bilirubin draws the universal model's AUROC was 0.277 (19 events), no better than in sparsely drawn patients (0.425), ruling out the possibility that denser sampling rescues the inversion and confirming construct non-transportability as the explanation. Administrative censoring was substantial: 48.8% of the development risk set and 35-59% of external risk sets left the ICU before completing the 72 h window; restricting the primary analysis to window-completers (n=2,349; 444 events) gave AUROC 0.789, a modest degradation consistent with diluted rather than inflated performance. Calibration within the diagnosis-matched eICU cirrhosis subgroup showed mean predicted risk 0.212 against an observed 10.2% event rate; a subgroup-specific intercept shift of -1.07 restored calibration-in-the-large and improved the Brier score from 0.097 to 0.078.

Label heterogeneity and clinical anchoring

Sepsis-3 onset within the preceding 48 h accompanied 39.9% of bilirubin-deterioration events in the MIMIC-IV cirrhosis cohort, 30.5% in SICdb, 20.6% in AmsterdamUMCdb, and 18.2% in eICU (Table 6), quantifying how strongly the laboratory surrogate overlaps with systemic inflammation and how much this overlap varies across health systems. In the

prespecified sensitivity analysis excluding sepsis-proximate events, the composite endpoint retained 290 of 613 events and the portable endpoint 112 of 240; discrimination declined modestly but remained useful (composite AUROC 0.817 to 0.753; portable 0.886 to 0.856; Table 9). Predicted risk was associated with death: hospital (or 30-day) mortality increased across portable-model risk quartiles in every external database – 6.6% to 17.4% in eICU, 11.4% to 21.8% in SICdb, 37.7% to 62.6% (30-day) in AmsterdamUMCdb, and 12.9% to 22.1% in NWICU (Table 7) – with logistic trend tests significant at $p < 0.001$ in the first three databases and positive but non-significant in NWICU ($p = 0.068$), whose small size (652 stays) produced non-significant middle-quartile dips. As noted in the Methods, these gradients cannot separate endpoint validity from the severity-marker properties of the constituent laboratory variables.

Discussion

In this multi-database study of 6,703 cirrhosis ICU stays and 63,268 external all-comer ICU stays across three countries, we rebuilt a landmark dynamic prediction framework for early hepatic deterioration and subjected it to the three tests that previous work in this population has avoided: benchmarking against naive rules, transportability assessment across health systems, and explicit quantification of outcome-label heterogeneity. Three findings matter. First, with a 21-fold larger event count than our previous single-centre effort, early laboratory dynamics predicted impending composite hepatic deterioration with an AUROC of 0.817 (95% CI 0.800-0.834) at the clinically actionable 24 h landmark / 48 h horizon – and, critically, exceeded the strongest of four tested naive rules – including a gap-completion rule mechanistically matched to the crossing definition – by +0.118 AUROC (95% CI 0.095-0.142), demonstrating genuine multi-analyte trajectory integration rather than threshold proximity masquerading as prediction. Second, a nine-analyte portable model discriminated consistently across four independent ICU databases (AUROC 0.876-0.901) but required local intercept recalibration, its raw outputs overestimating risk four- to six-fold in lower-prevalence all-comer populations. Third, 18-40% of bilirubin-deterioration events followed sepsis onset within 48 h, and excluding these events degraded discrimination only modestly – converting the surrogate-purity concern from a matter of assertion into a measured, system-dependent quantity.

The contrast with our previous report (an unpublished single-centre study) is instructive and illustrates the perils of small validation-free prediction studies. The earlier HBV-cirrhosis

cohort (392 stays; 46 events) yielded a point estimate of 0.725 with a confidence interval spanning 0.619-0.819, monitoring-intensity features dominated importance, and no external assessment existed. In the present development cohort the same framework, applied to 971 events, produced tighter inference, materially better discrimination, an aetiologically stable performance (the HBV subgroup event rate was 10.9% over the full stay and 14.3% at the primary combination, versus 11.7% in the original 392-stay study), and a sharply attenuated monitoring-bias signature (0.817 to 0.813 after removing monitoring variables, versus 0.727 to 0.655 previously). The earlier fragility most plausibly reflected event starvation rather than an intrinsic property of laboratory-based prediction, although residual monitoring information – embedded in the mere availability of serial measurements – cannot be excluded.

The naive-benchmark results define the honest boundaries of the contribution. For the composite endpoint, even the mechanistically matched gap-completion rule reached only 0.699 and the model's +0.118 increment over it – and +0.15 to +0.23 over every other rule – is substantial by any standard: integrating trajectories across four liver analytes plus coagulation, renal, and haematological markers does something no bedside rule can. For the single-analyte portable endpoint, however, the last bilirubin value alone reached 0.864 and the model added +0.032 – useful but modest. We report this asymmetry deliberately: the portable model should be understood as an efficiently packaged bilirubin-trajectory tracker whose cross-database discrimination is the finding of interest, not as evidence that machine learning extracts hidden signal from a single analyte.

Our external assessment combined transportability analysis with a diagnosis-matched validation wherever raw diagnosis tables could be analysed [27]. In eICU – the only database with both adequate laboratory coverage and enough diagnosis-matched cirrhosis patients – the conditional portable model retained an AUROC of 0.808 (0.748-0.865) among 591 coded-cirrhosis patients, strengthening to 0.848 (0.790-0.897) under the strict keyword variant; the deficit relative to the non-cirrhosis 0.896 was statistically significant (difference -0.088, 95% CI -0.151 to -0.035) and of the magnitude expected from case-mix narrowing, converting the central claim of this study into an aetiology-matched result. The multivariable cirrhosis-probability phenotype proved valid against coded diagnoses externally (AUROC 0.888-0.965; 0.816-0.912 without any liver test), anchoring the phenotype stratification used in databases where diagnosis tables remain unavailable, notably AmsterdamUMCdb. The diagnosis-matched analysis also delivered the study's most instructive negative finding, and

the robustness analyses sharpened it. The universal worsening endpoint inverted within eICU cirrhosis patients (AUROC 0.414-0.439) and was not rescued among densely sampled patients (0.277 with three or more post-24 h bilirubin draws), while its development label was predictable from raw draw count alone at AUROC 0.807. Taken together, the universal family is doubly limited: its label is partly a measurement-frequency construct, and the construct itself – dominated by doubling-based progression in the densely monitored development cohort but by threshold-crossing among detectable events in eICU – does not transport across measurement cultures. We therefore present the universal endpoints as hypothesis-generating for systems with at least daily liver-test monitoring, and the conditional threshold-crossing endpoint, whose single-crossing label logic is robust to sparse sampling and whose diagnosis-matched performance held at 0.808-0.848, as the deployable member of the family.

The sepsis co-occurrence analysis gives the surrogate-endpoint debate an empirical footing. Between 18% and 40% of bilirubin-deterioration events followed a Sepsis-3 onset within 48 h [26], with the highest overlap in the MIMIC-IV cirrhosis cohort; excluding these events left AUROCs of 0.753 (composite) and 0.856 (portable). This does not invalidate the endpoint: in an ICU screening context, a bilirubin trajectory about to cross the injury threshold warrants attention regardless of whether the proximate driver is sepsis-associated cholestasis, drug injury, or ischaemic hepatitis. It does define the correct interpretation of model output – a prompt for intensified hepatic monitoring, medication review, and diagnostic evaluation, not a probabilistic diagnosis of DILI [12,13]. The mortality gradients across predicted-risk quartiles in all four systems are consistent with this screening role, though they cannot, by construction, distinguish endpoint validity from the severity-marker properties of bilirubin, creatinine, and lactate.

Consistent with our previous findings, ICU-recorded hepatotoxic drug exposure again added no incremental discrimination (AUROC 0.511 alone; no gain in combination). This negative finding is bounded by the data: administration-record variables are crude proxies without dose, duration, or cumulative exposure, oral outpatient hepatotoxins are not captured, and their information is plausibly absorbed by the severity and laboratory variables that drive prescribing. Improving DILI-specific prediction will require dose-level exposure characterisation rather than more elaborate models of binary exposure flags.

This study has limitations that qualify its conclusions. First, both endpoint families are

laboratory-based surrogates without physician adjudication; we quantify their sepsis overlap and anchor them to mortality, but causal attribution to drug injury is impossible by construction. Second, the universal worsening endpoint is presented as hypothesis-generating: its progression criterion is one choice among several severity anchors, its development label is partly a measurement-frequency construct (draw count alone AUROC 0.807), and its construct did not transport to eICU cirrhosis patients even under dense sampling. Third, diagnosis-matched validation was powered only in eICU; SICdb records only primary admission diagnoses (73 flagged cases) and NWICU combines sparse laboratory coverage with small cirrhosis numbers, while AmsterdamUMCdb's diagnosis table was unavailable, so the Dutch validation remains phenotype-based. Fourth, external populations are predominantly all-comer cohorts with lower event rates, and laboratory sampling density differs several-fold across systems, affecting endpoint ascertainment in database-specific ways. Fifth, administrative censoring was substantial (35-59% of at-risk stays discharged before completing the 72 h window); event rates are therefore underestimates, and although the window-completers analysis (AUROC 0.789 vs 0.817) indicates diluted rather than inflated discrimination, competing-risk formulations are a preferable next step. Sixth, the AmsterdamUMCdb mortality anchor uses 30-day rather than hospital mortality because hospital-mortality flags were unavailable. Seventh, the primary-combination designation, although carried from the design of our previous study, was not externally preregistered, and with 60-plus model-endpoint-combination cells reported, the full tables should be read descriptively. Eighth, the Platt recalibration and external intercept correction are fitted and evaluated on the same data and are therefore optimistically biased; leave-one-database-out recalibration checks are planned for the prospective phase. Finally, all databases originate from high-income Western systems; transportability to high HBV-prevalence regions remains untested, and the development cohort's HBV subgroup (3% of stays), although performing comparably to other aetiologies, remains too small for standalone modelling.

In conclusion, a multi-landmark framework built on updated MIMIC-IV data demonstrates that early laboratory dynamics predict impending hepatic deterioration in ICU patients with cirrhosis with clinically useful, aetiologically stable discrimination that clearly exceeds naive threshold-proximity rules; that a parsimonious nine-analyte conditional model transports across four international ICU databases and retains discrimination within diagnosis-matched cirrhosis patients in eICU at a development-consistent event rate; that a universal worsening

endpoint extends coverage to patients already meeting laboratory injury criteria but is deployment-safe only where laboratory monitoring is at least daily; and that a validated cirrhosis-probability phenotype extends population-restricted assessment to databases without diagnosis tables. The findings reposition laboratory-trajectory early warning from a single-centre promise to a reproducible, multi-system signal suitable for prospective evaluation as a screening trigger for intensified hepatic surveillance, complementing organ-failure-based prognostic instruments in this population [28].

List of abbreviations

ALP: alkaline phosphatase; ALT: alanine aminotransferase; aPTT: activated partial thromboplastin time; AST: aspartate aminotransferase; AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision-recall curve; CI: confidence interval; DCA: decision curve analysis; DILI: drug-induced liver injury; HBV: hepatitis B virus; HCV: hepatitis C virus; ICD: International Classification of Diseases; ICU: intensive care unit; INR: international normalised ratio; IQR: interquartile range; LR: logistic regression; MELD: Model for End-Stage Liver Disease; OASIS: Oxford Acute Severity of Illness Score; PPV: positive predictive value; PT: prothrombin time; SAPS: Simplified Acute Physiology Score; SHAP: SHapley Additive exPlanations; SOFA: Sequential Organ Failure Assessment; TBIL: total bilirubin; TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

Declarations

Ethics approval and consent to participate: All databases are publicly available de-identified resources released under credentialed data use agreements. The MIMIC-IV programme was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology, which waived the requirement for individual informed consent; the remaining source databases were released under equivalent approvals from their institutions. No direct participant contact occurred, and no additional institutional review was required for this analysis of previously collected de-identified data.

Consent for publication: Not applicable.

Availability of data and materials: All source databases are available to qualified researchers after completion of the required training and data use agreements (PhysioNet for MIMIC-IV, eICU-CRD, SICdb, and NWICU; AmsterdamUMCdb via its institutional portal). The

harmonisation conventions used for this analysis (unit normalisation, outlier handling, stay-window assignment) are documented in Additional file 1; the completed TRIPOD+AI checklist is provided as Additional file 2; all analysis code are available from the corresponding author.

Funding: This research received no external funding.

Competing interests: The authors declare no competing interests.

Authors' contributions: Conceptualisation: ZZhiZ, HG; Methodology: ZZhiZ, ZZhenZ; Software and formal analysis: ZZhiZ; Investigation: ZZhiZ, ZZhenZ; Data curation: ZZhiZ; Writing – original draft: ZZhiZ; Writing – review and editing: ZZhiZ, ZZhenZ, HG; Supervision: HG. All authors read and approved the final manuscript.

Acknowledgements: The authors thank the institutions and investigators who created and maintain the MIMIC-IV, eICU-CRD, AmsterdamUMCdb, SICdb, and NWICU resources.

References

1. Ledgerwood C, Villgran V, Mardirossian N, et al. Acute liver failure. *Crit Care Nurs Q*. 2022;45(3):248-257. doi:10.1097/CNQ.0000000000000409
2. Stravitz RT, Fontana RJ, Karvellas CJ, et al. Future directions in acute liver failure. *Hepatology*. 2023;78(4):1276-1289.
3. Perez Ruiz de Garibay A, Kortgen A, Leonhardt J, et al. Critical care hepatology: definitions, incidence, prognostic value and clinical implications. *Crit Care*. 2022;26:293.
4. Moreau R, Jalan R, Gines P, et al. Acute-on-chronic liver failure is a distinct syndrome that develops in patients with cirrhosis. *Gastroenterology*. 2013;144(5):1426-1437.
5. Bajaj JS, O'Leary JG, Reddy KR, et al. Survival in infection-related acute-on-chronic liver failure is defined by extrahepatic organ failures. *Hepatology*. 2014;60(1):250-256.
6. van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *J Am Stat Assoc*. 2007;102(478):528-537.
7. Lavalley-Morelle A, Timsit JF, Mentre F, et al. Joint modeling under competing risks: application to survival prediction in patients admitted in intensive care unit for sepsis. *CPT Pharmacometrics Syst Pharmacol*. 2022;11(11):1472-1484.
8. Wei J, Lin F, Jin T, et al. Dynamic SOFA-based deep learning for mortality prediction in sepsis survivors. *J Big Data*. 2025;12:76.
9. Tan Y, Dede M, Mohanty V, et al. Forecasting acute kidney injury and resource utilization in the intensive care unit. *J Biomed Inform*. 2024;153:104635.
10. Zhang J, Bolanos Trujillo LD, Tanwar A, et al. Clinical utility of automatic phenotype annotation in the intensive care unit. *BMJ Health Care Inform*. 2022;29(1):e100460.
11. Chen M, Suzuki A, Thakkar S, Yu K, Hu C, Tong W. DILrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov Today*. 2016;21(4):648-653.
12. Andrade RJ, Chalasani N, Bjornsson ES, et al. Drug-induced liver injury. *Nat Rev Dis Primers*.

2019;5(1):58.

13. Chalasani NP, Hayashi PH, Bonkovsky HL, et al. ACG Clinical Guideline: Diagnosis and management of drug-induced liver injury. *Am J Gastroenterol*. 2021;116(5):878-898.
14. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10:1.
15. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data*. 2018;5:180178.
16. Thorat PP, Peelen LM, Snijders KE, et al. Sharing patient-level data from observational studies in intensive care: advantages and challenges of the AmsterdamUMCdb cohort database. *Crit Care Med*. 2021;49(9):e871-e877.
17. Rodemund N, Kokoefer A, Wernly B, Cozowicz C. The Salzburg Intensive Care database (SICdb), a freely accessible intensive care database. *Sci Data*. 2023;10:4. doi:10.1038/s41597-022-01899-x
18. Northwestern ICU (NWICU) Database (version 0.1.0). PhysioNet. <https://physionet.org/content/nwicu-northwestern-icu/>. Accessed 26 Aug 2026.
19. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
20. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
21. Kamath PS, Wiesner RH, Malinchoc M, et al. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology*. 2001;33(2):464-470.
22. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3146-3154.
23. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.
24. Vickers AJ, van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, diagnostic tests, and biomarkers. *Nat Med*. 2019;35:1678-1684.
25. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765-4774.
26. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*. 2016;315(8):801-810.
27. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58.
28. Jalan R, Saliba F, Pavesi M, et al. Development and validation of a prognostic score to predict mortality in patients with acute-on-chronic liver failure. *J Hepatol*. 2014;61(5):1038-1047.

Tables

Table 1. Development cohort characteristics (MIMIC-IV 3.1, cirrhosis ICU stays, n = 6,703).

Characteristic	Value
Age, years, median (IQR)	60 (52-67)
Male, n (%)	4,320 (64.4)

Characteristic	Value
ICU length of stay, days, median (IQR)	2.3 (1.2-4.9)
MELD, median (IQR)	23 (15-31)
First-day SOFA, median (IQR)	7 (5-10)
SAPS-II, median (IQR)	38 (29-49)
OASIS, median (IQR)	31 (26-38)
Charlson index, median (IQR)	6 (4-8)
Aetiology: alcohol / HCV / HBV, n (%)	3,603 (53.8) / 831 (12.4) / 201 (3.0)
Hospital mortality, n (%)	1,639 (24.5)
Gradable baseline (endpoint A), n	5,509
Composite hepatic deterioration (soft), n (%)	971 (17.6)
Composite hepatic deterioration (strict), n (%)	567 (10.3)

Table 2. Development performance (out-of-fold, GroupKFold by patient) across landmark-horizon combinations, endpoint A (composite hepatic deterioration), LightGBM versus logistic regression.

Landmark (h)	Horizon (h)	n	Events	LGBM AUROC	LGBM AUPRC	LR AUROC
6	24	5,389	190	0.750	0.161	0.666
6	48	5,389	513	0.746	0.257	0.653
6	72	5,389	632	0.729	0.282	0.649
12	24	5,296	343	0.792	0.254	0.689
12	48	5,296	567	0.774	0.335	0.687
12	72	5,296	656	0.751	0.337	0.677
24	24	4,585	462	0.845	0.466	0.701
24	48 (primary)	4,585	613	0.817	0.478	0.678
24	72	4,585	697	0.796	0.462	0.687
48	24	2,835	151	0.651	0.109	0.519
48	48	2,835	235	0.682	0.173	0.567
48	72	2,835	296	0.712	0.243	0.571

Primary combination bootstrap 95% CI: AUROC 0.800-0.834; AUPRC 0.437-0.521. The 24 h / 48 h designation was fixed before analysis, carried from the design of our previous study. Endpoint B (portable bilirubin endpoint) development: AUROC 0.886, AUPRC 0.564 (n = 1,991; 240 events; 12.1%).

Table 3. External transportability of the portable model (LightGBM, endpoint B, 24 h landmark / 48 h horizon), applied without refitting; bootstrap 95% CIs.

Database (country)	Population	n	Events (rate)	AUROC (95% CI)	AUPRC
eICU (US)	All gradable	50,496	970 (1.9%)	0.897 (0.887-0.907)	0.218
eICU (US)	Liver-vulnerable	15,115	600 (4.0%)	0.870 (0.854-0.884)	0.267
eICU (US)	Non-vulnerable	35,381	370 (1.0%)	0.884 (0.865-0.901)	0.134
SICdb (Austria)	All gradable	5,455	168 (3.1%)	0.880 (0.847-0.908)	0.334
SICdb (Austria)	Liver-vulnerable	1,616	108 (6.7%)	0.884 (0.851-0.914)	0.423
SICdb (Austria)	Non-vulnerable	3,839	60 (1.6%)	0.821 (0.757-0.877)	0.170

Database (country)	Population	n	Events (rate)	AUROC (95% CI)	AUPRC
AmsterdamUMCdb (NL)	All gradable	6,665	135 (2.0%)	0.876 (0.841-0.906)	0.209
AmsterdamUMCdb (NL)	Liver-vulnerable	2,443	102 (4.2%)	0.835 (0.796-0.873)	0.224
AmsterdamUMCdb (NL)	Non-vulnerable	4,222	33 (0.8%)	0.841 (0.751-0.918)	0.189
NWICU (US)	All gradable	652	12 (1.8%)	0.901 (0.802-0.976)	0.367
NWICU (US)	Liver-vulnerable	207	6 (2.9%)	0.952 (0.899-0.996)	0.540

Liver-vulnerable phenotype: minimum first-24 h platelet count $<150 \times 10^9/L$ (internal validation vs ICD-coded cirrhosis: AUROC 0.798, sensitivity 0.773, specificity 0.685, PPV 0.296). Calibration: raw mean predicted risk exceeded observed rates 4-6 fold; intercept-only correction shifts of -1.79 to -2.39 restored calibration-in-the-large and improved Brier scores in all databases (e.g., eICU 0.026 to 0.017). NWICU intervals are wide owing to 12 events. Logistic calibration slopes after intercept correction (which does not alter the slope): eICU 1.23, SICdb 1.01, AmsterdamUMCdb 1.06, NWICU 1.32.

Table 4. Ablation analysis at the primary combination (endpoint A, 24 h / 48 h, LightGBM).

Feature set	AUROC	AUPRC
Static (age, sex)	0.508	0.139
Severity scores	0.667	0.225
Laboratory dynamics	0.816	0.479
Laboratory dynamics without monitoring-intensity variables	0.813	0.458
Hepatotoxic medication exposure (ICU administration records)	0.511	0.142
Static + scores	0.667	0.223
Static + laboratory	0.816	0.479
Static + scores + laboratory	0.817	0.477
Full model	0.817	0.478

Table 5. Primary model performance by cirrhosis aetiology (endpoint A, 24 h / 48 h), with bootstrap 95% CIs.

Subgroup	n	Events	AUROC (95% CI)
HBV-related	119	17	0.813 (0.698-0.914)
HCV-related	578	83	0.848 (0.801-0.894)
Alcohol-related	2,576	323	0.825 (0.797-0.851)

Table 6. Sepsis-3 onset within 48 h before bilirubin-deterioration events (endpoint B), by database.

Database	Events	With preceding sepsis, n (%)
MIMIC-IV cirrhosis cohort	328	131 (39.9%)
SICdb	374	114 (30.5%)
AmsterdamUMCdb	311	64 (20.6%)
eICU	1,842	335 (18.2%)

Table 7. Mortality by portable-model predicted-risk quartile in external databases (quartile 1 = lowest risk; logistic trend test p-values shown).

Database	Q1	Q2	Q3	Q4
eICU, hospital mortality (p<0.001)	6.6%	8.3%	11.0%	17.4%
SICdb, hospital mortality (p<0.001)	11.4%	14.7%	14.6%	21.8%
AmsterdamUMCdb, 30-day mortality (p<0.001)	37.7%	50.8%	53.5%	62.6%
NWICU, hospital mortality (p=0.068)	12.9%	14.7%	9.8%	22.1%

Table 8. Naive-rule benchmarks versus models at the primary combination (24 h / 48 h).

Endpoint / benchmark	AUROC
Endpoint A – naive: threshold proximity (max last/threshold)	0.585
Endpoint A – naive: relative rise (max delta/threshold)	0.661
Endpoint A – naive: proximity + rise	0.603
Endpoint A – naive: gap-completion (fraction of baseline-to-threshold distance covered)	0.699
Endpoint A – LightGBM model	0.817
Endpoint A – incremental vs best naive (gap-completion)	+0.118 (0.095 to 0.142), p<0.001
Endpoint B – naive: last TBIL	0.864
Endpoint B – naive: delta TBIL	0.680
Endpoint B – naive: last + delta TBIL	0.854
Endpoint B – fixed rule (last >=1.8 or delta >=0.3 mg/dL)	sens 0.742, spec 0.805, PPV 0.343
Endpoint B – LightGBM model	0.886
Endpoint B – incremental vs best naive	+0.032 (0.014 to 0.052), p=0.002

Table 9. Sensitivity analysis excluding events with Sepsis-3 onset within the preceding 48 h (primary combination, 24 h / 48 h).

Endpoint	Analysis	n	Events	AUROC	AUPRC
A (composite)	Original	4,585	613	0.817	0.478
A (composite)	Sepsis-proximate events excluded	4,262	290	0.753	0.208
B (portable)	Original	1,991	240	0.886	0.564
B (portable)	Sepsis-proximate events excluded	1,863	112	0.856	0.317

Table 10. Universal worsening endpoint (covering already-abnormal baselines): development performance and external transportability (LightGBM, 24 h landmark / 48 h horizon unless stated).

Analysis	n	Events (rate)	AUROC (95% CI)	AUPRC
A2 composite – development	4,671	2,674 (57.2%)	0.895 (0.884-0.905)	0.908
A2 composite – naive proximity+rise	4,671	-	0.861	-
A2 composite – incremental vs naive	-	-	+0.033 (0.026-0.041)	-
B2 portable – development	4,648	2,263 (48.7%)	0.914 (0.906-0.922)	0.891
B2 portable – naive last TBIL	4,648	-	0.892	-
B2 portable – incremental vs naive	-	-	+0.050 (0.042-0.059)	-
B2 portable – eICU external	55,351	1,104 (2.0%)	0.844 (0.835-0.852)	0.064
B2 portable – SICdb external	6,067	193 (3.2%)	0.825 (0.796-0.851)	0.104

Analysis	n	Events (rate)	AUROC (95% CI)	AUPRC
B2 portable – AmsterdamUMCdb external	7,072	152 (2.1%)	0.853 (0.824-0.880)	0.110
B2 portable – NWICU external	719	12 (1.7%)	0.814 (0.675-0.911)	0.060

Universal event = new threshold crossing OR (baseline at/above threshold) peak $\geq 2 \times$ baseline with increment ≥ 40 U/L (ALT/AST), ≥ 1.0 mg/dL (TBIL), ≥ 80 U/L (ALP). All-component events in development: progression 2,503, new crossing 971 of 2,827 stays with any event. Other landmark-horizon combinations for A2: AUROC 0.818-0.860 (landmarks 6-12 h) and 0.815-0.834 (landmark 48 h).

Table 11. Cirrhosis-probability phenotype: internal validation and phenotype-restricted external validation of the universal portable model (B2).

Analysis	Metric	Value
Phenotype, full (9 labs + age/sex), internal	AUROC vs ICD cirrhosis (n=45,621; 6,620 cases)	0.910 (0.906-0.915)
Phenotype, restricted (no liver tests), internal	AUROC vs ICD cirrhosis	0.868 (0.862-0.872)
eICU, top 30% phenotype probability	Event rate / model AUROC	5.2% / 0.646
eICU, remaining 70%	Event rate	0.6%
eICU, top 30% restricted phenotype	Event rate / model AUROC	4.0% / 0.748
SICdb, top 30% phenotype probability	Event rate / model AUROC	7.9% / 0.655
SICdb, remaining 70%	Event rate	1.2%
SICdb, top 30% restricted phenotype	Event rate / model AUROC	6.6% / 0.763
AmsterdamUMCdb, top 30% phenotype probability	Event rate / model AUROC	5.7% / 0.748
AmsterdamUMCdb, remaining 70%	Event rate	0.6%
AmsterdamUMCdb, top 30% restricted phenotype	Event rate / model AUROC	4.9% / 0.786

Event = universal bilirubin deterioration endpoint (B2). Within-stratum AUROC is attenuated relative to all-comer values because high-probability strata are clinically homogeneous; the restricted phenotype avoids circularity with liver-focused endpoints at a modest cost in enrichment.

Table 12. True diagnosis-matched external validation using raw diagnosis tables (cirrhosis: ICD-9 571.2/571.5/571.6, ICD-10 K70.2-K70.4 and K74.x, plus text keywords).

Database / analysis	n	Events (rate)	AUROC (95% CI)
eICU, endpoint B, cirrhosis	591	60 (10.2%)	0.808 (0.748-0.865)
eICU, endpoint B, non-cirrhosis	49,905	910 (1.8%)	0.896 (0.886-0.906)
eICU, endpoint B2, cirrhosis	1,512	83 (5.5%)	0.414 (0.369-0.461)
eICU, endpoint B2, non-cirrhosis	53,839	1,021 (1.9%)	0.854 (0.845-0.863)
SICdb, cirrhosis (B / B2)	7 / 30	1 / 2	too small for estimation
NWICU, cirrhosis (B / B2)	6 / 22	2 / 2	too small for estimation
Phenotype vs diagnoses, eICU (full / restricted)	55,351	1,512 cirrhosis	0.888 (0.880-0.897) / 0.816 (0.804-0.826)
Phenotype vs diagnoses, SICdb (full / restricted)	6,067	30 cirrhosis	0.965 (0.946-0.979) / 0.912 (0.879-0.942)

Database / analysis	n	Events (rate)	AUROC (95% CI)
Phenotype vs diagnoses, NWICU (full / restricted)	719	22 cirrhosis	0.961 (0.935-0.982) / 0.909 (0.849-0.957)

AmsterdamUMCdb's diagnosis table was unavailable; that database remains validated through the phenotype approach. The universal-endpoint inversion within eICU cirrhosis reflects sparse bilirubin sampling (doubling-based progression largely undetectable); see Results.

Table 13. Second-round robustness analyses.

Analysis	Result
Universal label vs bilirubin draw count alone (B2, development)	AUROC 0.807 (0.794-0.819)
Universal composite label vs liver-panel draw count (A2)	AUROC 0.867
B2 without monitoring-intensity features	0.914 to 0.906
A2 without monitoring-intensity features	0.895 to 0.889
Trajectory-extrapolation naive rule (A2 / B2)	0.574 / 0.607 (model increments +0.321 / +0.307)
eICU cirrhosis, strict keywords, endpoint B	n=471, 43 events, AUROC 0.848 (0.790-0.897)
eICU cirrhosis, strict keywords, endpoint B2	n=1,162, 59 events, AUROC 0.439 (0.393-0.490)
eICU cirrhosis B2, >=3 post-24 h bilirubin draws	n=128, 19 events, AUROC 0.277
Cirrhosis minus non-cirrhosis AUROC (eICU, endpoint B)	-0.088 (-0.151 to -0.035), p<0.001
Cirrhosis-subgroup calibration (eICU, endpoint B)	predicted 0.212 vs observed 0.102; shift -1.07; Brier 0.097 to 0.078
Censored before 72 h window completion	development 48.8%; external 35-59%
Primary model in window-completers only	n=2,349, 444 events, AUROC 0.789

Figure legends

Figure 1. Out-of-fold AUROC across the 12 landmark-horizon combinations (LightGBM). (A) Endpoint A, composite hepatic deterioration, MIMIC-IV cirrhosis cohort. (B) Endpoint B, portable bilirubin deterioration. Discrimination peaks at the 24 h landmark in both endpoints.

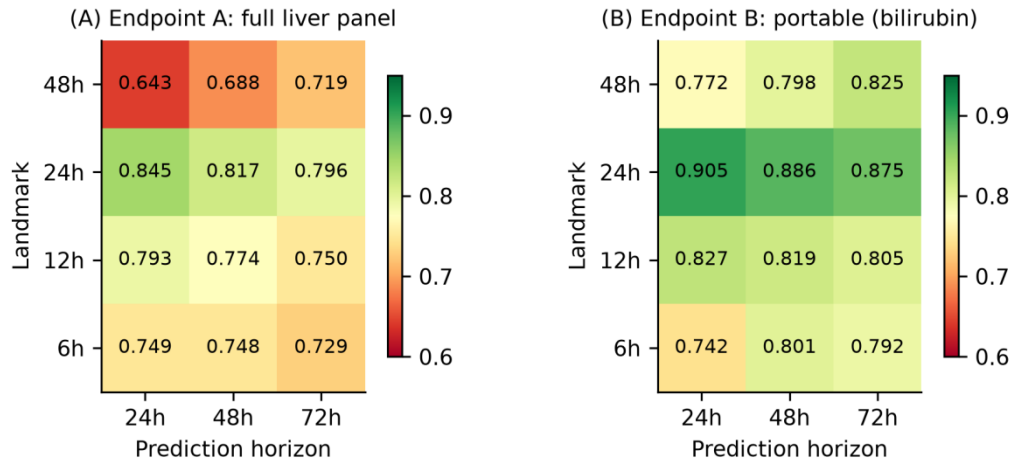


Figure 2. Primary model (endpoint A, 24 h landmark / 48 h horizon). (A) ROC curve; (B) precision-recall curve against the 13.4% event rate; (C) calibration plot with Platt recalibration; (D) decision curve analysis on recalibrated probabilities (raw-probability curve shown for comparison; net benefit of raw outputs is distorted by the known class-weighting overdispersion).

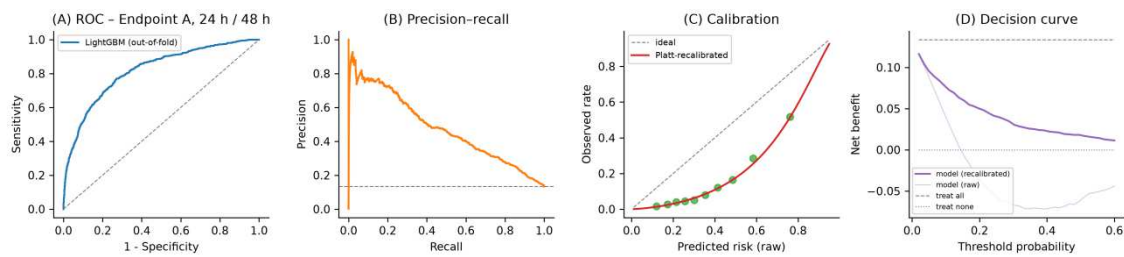


Figure 3. External transportability of the portable model. (A) ROC curves in the four external databases; (B) calibration deciles. The model was applied without refitting; raw predicted risks exceed observed event rates four- to six-fold and are restored by a single intercept correction.

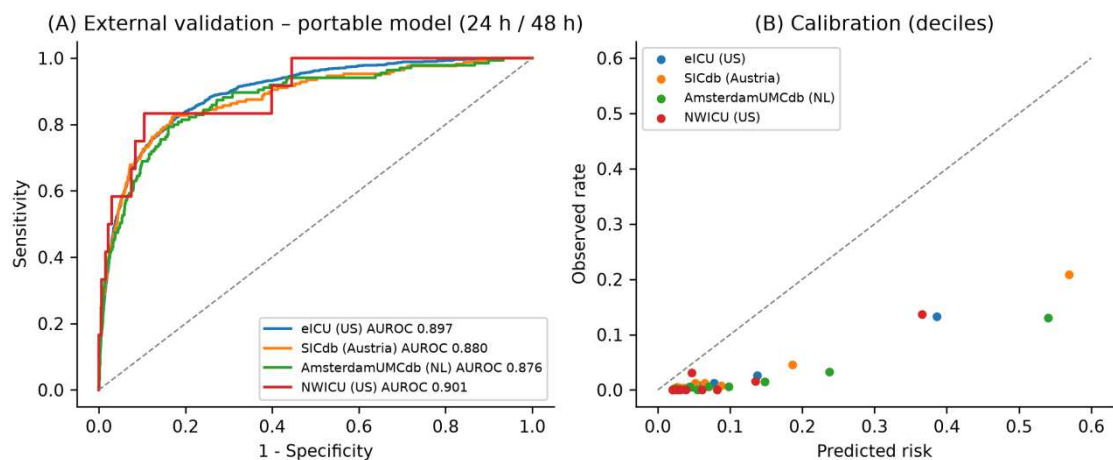
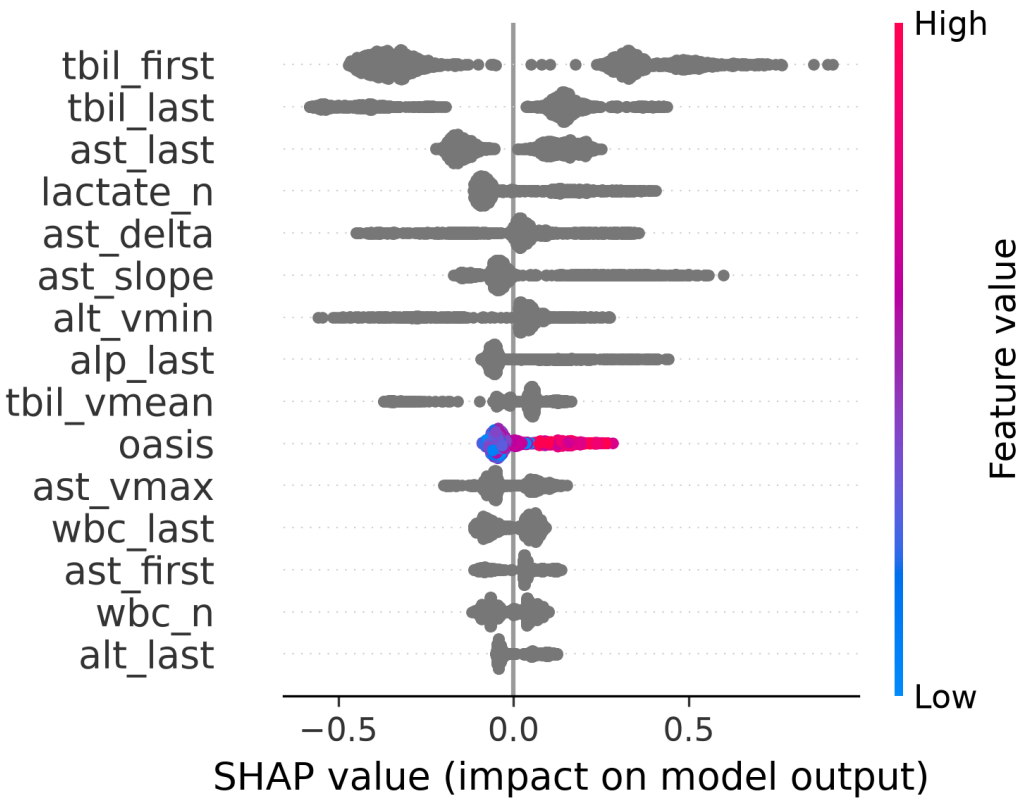


Figure 4. SHAP summary plot for the primary model (15 most influential features).

Laboratory dynamics – particularly bilirubin, INR, AST and creatinine trajectories and their extremes – dominate the model.



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [03additionalfile1supplement.docx](#)
- [04additionalfile2TRIPODAIchecklist.docx](#)