

Neonatal Bilirubin Trajectory Phenotypes and Composite Adverse Outcomes: A Two-Database Latent-Class Analysis (MIMIC-III and PIC)

Jingchuan Lu

ljcspu@leshan-hospital.org.cn

Leshan People's Hospital

Meiding Wang

Leshan People's Hospital

Haiyi Xu

Leshan People's Hospital

Shuangyan Zhu

Leshan People's Hospital

Article

Keywords: bilirubin, trajectory, latent class, neonate, composite outcome, risk stratification

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10886732/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Bilirubin is measured routinely in neonates, but whether its trajectory over the first postnatal week adds risk-stratification value beyond a single measurement is unclear. In this preregistered two-database retrospective cohort study (MIMIC-III, $n = 8,059$; PIC, $n = 2,953$ neonates admitted at < 28 days of age), we estimated latent-class trajectories of log-transformed total bilirubin on postnatal days 0–7 (neonates with ≥ 3 measurements; $n = 2,951$ and $1,817$) and related them to a composite of in-hospital death, necrotizing enterocolitis, or hospital-acquired late-onset sepsis. Four phenotypes were identified (typical 62.6%, delayed high peak 9.4%, persistently low 6.8%, rapid-rise early peak 21.2%). The rapid-rise early-peak class had higher composite-outcome risk in both databases (OR 2.65, MIMIC-III; 1.90, PIC), attenuated to non-significance after birth-weight adjustment in MIMIC-III. The persistently low class had the lowest bilirubin values but the highest risk (OR 7.45; 2.55), largely explained by birth weight. The trajectory model improved discrimination over the Bhutani risk zone in MIMIC-III (corrected AUC 0.820 vs 0.768) but matched the peak value (0.820 vs 0.822); discrimination in PIC was modest. Trajectory phenotypes identified high-risk neonates consistently, yet incremental value over single-point measures was modest and associations partly reflected immaturity and severity.

Introduction

Neonatal jaundice is among the most common conditions in the neonatal period, and total serum bilirubin (TSB) is measured routinely in hospitalized neonates for the surveillance and management of hyperbilirubinemia [1–3]. The hour-specific Bhutani nomogram, which assigns a single TSB value to a postnatal-age-specific risk zone, has become the standard framework for assessing the risk of subsequent significant hyperbilirubinemia in infants born at ≥ 35 weeks of gestation [1, 2]. Beyond jaundice itself, bilirubin is biologically active: at low concentrations it is a potent antioxidant and may protect against oxygen-radical-mediated injury in preterm infants [4–6], whereas at high concentrations it is neurotoxic [7], and systemic inflammation and sepsis can impair bilirubin clearance, producing cholestasis that tracks illness severity [8, 9]. Consistent with a prognostic role beyond jaundice, higher TSB at NICU admission was associated with a lower risk of acute kidney injury in a large MIMIC-III cohort [10].

A single measurement, however, captures the bilirubin level at one time point only and discards its temporal shape. Physiologically, TSB follows a characteristic postnatal course - rising after birth, peaking, and then declining - and pathological processes (e.g., hemolysis, cholestasis, or sepsis-related hepatic dysfunction) may alter both the height and the timing of this curve [7, 8, 9]. Whether the shape of the early bilirubin trajectory, rather than any single value, adds risk-stratification value is unknown.

Latent-class trajectory modeling provides a principled framework for identifying subgroups that share a similar temporal course of a longitudinal biomarker [11, 12] and trajectory-based methods are increasingly used for the analysis and reporting of longitudinal biomarker trajectories [13]; trajectory approaches have been applied to neonatal outcomes including bronchopulmonary dysplasia,

neurodevelopment, and growth [14]. To our knowledge, no study has characterized latent-class trajectories of early bilirubin in relation to composite neonatal adverse outcomes, and previous single-point analyses were single-center [10].

We therefore conducted a preregistered, two-database retrospective cohort study (MIMIC-III, USA; PIC, China) with two objectives: (i) to identify latent-class trajectories of TSB during postnatal days 0–7; and (ii) to determine whether trajectory phenotype membership improves risk stratification for a composite outcome of in-hospital death, necrotizing enterocolitis (NEC), or hospital-acquired late-onset sepsis (LOS) over three single-point measures (first value, peak value, and Bhutani risk zone [1, 2, 10]).

Results

Cohort and baseline characteristics

The full cohort comprised 8,059 (MIMIC-III) and 2,953 (PIC) neonates admitted before 28 days of age; the trajectory analysis included 2,951 and 1,817 neonates with at least three TSB measurements on postnatal days 0–7, respectively (Fig. 1). Four latent-class trajectories were identified (Fig. 2): a typical/reference course (C2; 58.9% of MIMIC-III and 68.7% of PIC), a rapid-rise early-peak course (C4; 27.7% and 10.7%), a persistently low course (C3; 8.8% and 3.4%) with the lowest first and peak TSB values of all classes in both databases (MIMIC-III median first 3.8 mg/dL, peak 4.9 mg/dL; PIC 2.8 and 3.8 mg/dL), and a delayed high-peak course (C1; 4.5% and 17.3%). Baseline characteristics differed markedly across classes (Table 1). In MIMIC-III, the C3 class comprised predominantly very premature, very-low-birth-weight infants (preterm 97.7%; median admission weight 0.8 kg; vasoactive drugs 26.1%), with the highest rates of the composite outcome (31.4%) and prolonged hospitalization (79.7%). In PIC, by contrast, the C3 class comprised mostly term, normal-weight infants (preterm 9.8%; median admission weight 2.7 kg) yet also had a high composite-outcome rate (29.5%), driven mainly by death (23.0%). These compositional differences are considered in the Discussion.

Table 1
Baseline characteristics by trajectory class

MIMIC-III (n = 2,951)					
Characteristic	C1 (n = 134)	C2 (n = 1738)	C3 (n = 261)	C4 (n = 818)	P
Male, %	61.9	56.0	45.6	51.8	0.002
Admitted < 24 h of age, %	73.9	95.7	100.0	100.0	< 0.001
Admitted 24–72 h of age, %	1.5	0.9	0.0	0.0	0.011
Admitted 72 h-7 d of age, %	24.6	3.3	0.0	0.0	< 0.001
Admitted > 7 d of age, %	0.0	0.0	0.0	0.0	-
Preterm, %	58.2	78.3	97.7	93.5	< 0.001
Low birth weight, %	0.7	3.9	16.1	8.8	< 0.001
Congenital anomaly, %	21.6	17.9	61.7	34.8	< 0.001
Early-onset sepsis, %	0.0	0.6	1.1	0.7	0.601
Community-acquired/present-on-admission bacteremia, %	0.0	0.1	0.0	0.1	0.926
Vasoactive drugs within 24 h, %	8.2	2.6	26.1	5.5	< 0.001
Composite outcome (death/NEC/LOS), %	2.2	2.8	31.4	9.4	< 0.001
Death, %	2.2	0.5	4.6	0.6	< 0.001
NEC, %	0.0	1.4	14.9	4.9	< 0.001
Hospital-acquired LOS, %	0.0	1.3	18.4	5.0	< 0.001
Prolonged hospitalization (> 28 d), %	6.0	13.8	79.7	46.3	< 0.001
Admission weight, kg, median (IQR)	2.6 (2.2–3)	2.1 (1.7–2.5)	0.8 (0.6–0.9)	1.4 (1.1–1.7)	< 0.001
First TSB, mg/dL, median (IQR)	4.9 (3.9–13)	6.2 (5.1–9.2)	3.8 (2.9–5.1)	5.7 (4.7–6.7)	< 0.001
Peak TSB, mg/dL, median (IQR)	9.4 (7.6–17)	9 (8–9.7)	4.9 (4–6.2)	7.9 (6.6–9)	< 0.001
<i>C1, delayed high peak; C2, typical (reference); C3, persistently low; C4, rapid-rise early peak. Values are percentages of the class unless otherwise stated; medians are shown with interquartile ranges. P values are from chi-square tests (categorical) or Kruskal-Wallis tests (continuous) across the four classes within each database; P is omitted for characteristics with no between-class variation. TSB, total serum bilirubin; NEC, necrotizing enterocolitis; LOS, late-onset sepsis.</i>					

MIMIC-III (n = 2,951)					
Characteristic	C1 (n = 134)	C2 (n = 1738)	C3 (n = 261)	C4 (n = 818)	P
No. of TSB measurements, median (IQR)	4 (4–5)	4 (3–5)	6 (5–7)	5 (4–6)	< 0.001
PIC (n = 1,817)					
Characteristic	C1 (n = 314)	C2 (n = 1248)	C3 (n = 61)	C4 (n = 194)	P
Male, %	61.8	62.2	50.8	54.6	0.080
Admitted < 24 h of age, %	73.2	59.5	65.6	59.8	< 0.001
Admitted 24–72 h of age, %	16.6	25.7	26.2	29.4	0.003
Admitted 72 h–7 d of age, %	10.2	14.7	8.2	10.8	0.063
Admitted > 7 d of age, %	0.0	0.0	0.0	0.0	-
Preterm, %	30.3	28.4	9.8	26.8	0.011
Low birth weight, %	0.6	0.4	0.0	1.5	0.221
Congenital anomaly, %	31.8	28.3	14.8	22.2	0.012
Early-onset sepsis, %	2.5	1.7	1.6	2.6	0.683
Community-acquired/present-on-admission bacteremia, %	3.5	1.7	1.6	0.0	0.033
Vasoactive drugs within 24 h, %	7.3	4.6	1.6	4.1	0.116
Composite outcome (death/NEC/LOS), %	20.7	14.8	29.5	24.2	< 0.001
Death, %	14.0	5.6	23.0	12.9	< 0.001
NEC, %	1.3	2.2	1.6	2.6	0.694
Hospital-acquired LOS, %	8.6	8.6	8.2	10.3	0.880
Prolonged hospitalization (> 28 d), %	35.0	34.3	24.6	33.5	0.453
Admission weight, kg, median (IQR)	2.3 (1.4–2.9)	2.1 (1.3–3)	2.7 (1.9–3.4)	2.4 (1.2–3.1)	0.472
<i>C1, delayed high peak; C2, typical (reference); C3, persistently low; C4, rapid-rise early peak. Values are percentages of the class unless otherwise stated; medians are shown with interquartile ranges. P values are from chi-square tests (categorical) or Kruskal-Wallis tests (continuous) across the four classes within each database; P is omitted for characteristics with no between-class variation. TSB, total serum bilirubin; NEC, necrotizing enterocolitis; LOS, late-onset sepsis.</i>					

MIMIC-III (n = 2,951)					
Characteristic	C1 (n = 134)	C2 (n = 1738)	C3 (n = 261)	C4 (n = 818)	P
First TSB, mg/dL, median (IQR)	2.2 (1.5–4.2)	4.2 (2.7–8)	2.8 (2–3.7)	5.2 (3.2–7.2)	< 0.001
Peak TSB, mg/dL, median (IQR)	12 (8.7–15.1)	9.9 (7.9–12.4)	3.8 (2.9–5)	7.9 (6.3–9.8)	< 0.001
No. of TSB measurements, median (IQR)	7 (5–10)	5 (4–7.2)	5 (4–10)	6 (4–9)	< 0.001
<i>C1, delayed high peak; C2, typical (reference); C3, persistently low; C4, rapid-rise early peak. Values are percentages of the class unless otherwise stated; medians are shown with interquartile ranges. P values are from chi-square tests (categorical) or Kruskal-Wallis tests (continuous) across the four classes within each database; P is omitted for characteristics with no between-class variation. TSB, total serum bilirubin; NEC, necrotizing enterocolitis; LOS, late-onset sepsis.</i>					

Primary outcome: composite adverse outcome by trajectory class

Relative to the typical/reference class (C2), the persistently low class (C3) and the rapid-rise early-peak class (C4) had significantly higher composite-outcome risk in both databases (MIMIC-III: C3 OR 7.45 [4.82–11.53], C4 OR 2.65 [1.80–3.89]; PIC: C3 OR 2.55 [1.42–4.57], C4 OR 1.90 [1.31–2.74]; all $p < 0.01$; Table 2, Fig. 3). The delayed high-peak class (C1) was not associated with the composite outcome in MIMIC-III (OR 0.78 [0.23–2.58]) but was associated with higher risk in PIC (OR 1.50 [1.09–2.07], $p = 0.013$), a cross-database inconsistency addressed in the Discussion. Firth penalized-logistic models yielded nearly identical estimates (e.g., MIMIC-III C3 OR 7.37 [4.79–11.41]), indicating that results were not inflated by sparse-data separation.

Table 2
Adjusted associations between trajectory class and outcomes

Outcome	Database	Trajectory class	OR (95% CI)	P
Composite (death/NEC/LOS)	MIMIC-III	C2: Typical (reference)	1.00 (reference)	-
Composite (death/NEC/LOS)	MIMIC-III	C1: Delayed high peak	0.78 (0.23–2.58)	0.679
Composite (death/NEC/LOS)	MIMIC-III	C3: Persistently low	7.45 (4.82–11.53)	< 0.001
Composite (death/NEC/LOS)	MIMIC-III	C4: Rapid-rise early peak	2.65 (1.80–3.89)	< 0.001
Composite (death/NEC/LOS)	PIC	C2: Typical (reference)	1.00 (reference)	-
Composite (death/NEC/LOS)	PIC	C1: Delayed high peak	1.50 (1.09–2.07)	0.013
Composite (death/NEC/LOS)	PIC	C3: Persistently low	2.55 (1.42–4.57)	0.002
Composite (death/NEC/LOS)	PIC	C4: Rapid-rise early peak	1.90 (1.31–2.74)	< 0.001
Prolonged hospitalization (> 28 d)	MIMIC-III	C2: Typical (reference)	1.00 (reference)	-
Prolonged hospitalization (> 28 d)	MIMIC-III	C1: Delayed high peak	0.38 (0.18–0.83)	0.015
Prolonged hospitalization (> 28 d)	MIMIC-III	C3: Persistently low	12.16 (8.44–17.51)	< 0.001
Prolonged hospitalization (> 28 d)	MIMIC-III	C4: Rapid-rise early peak	4.01 (3.25–4.93)	< 0.001
Prolonged hospitalization (> 28 d)	PIC	C2: Typical (reference)	1.00 (reference)	-
Prolonged hospitalization (> 28 d)	PIC	C1: Delayed high peak	0.99 (0.74–1.33)	0.956
Prolonged hospitalization (> 28 d)	PIC	C3: Persistently low	0.89 (0.47–1.68)	0.725
Prolonged hospitalization (> 28 d)	PIC	C4: Rapid-rise early peak	0.98 (0.68–1.40)	0.899

OR, odds ratio; CI, confidence interval; NEC, necrotizing enterocolitis; LOS, late-onset sepsis. Models were adjusted for sex; postnatal age at admission; ICD-based preterm birth and low birth weight; congenital anomalies; early-onset sepsis; community-acquired/present-on-admission bacteremia; and vasoactive drug use within the first 24 h. The typical class (C2) served as the reference. Primary-analysis estimates for the composite outcome are shown in Fig. 3.

Component-wise analysis

In MIMIC-III, the higher composite risk of C3 and C4 was driven by NEC (C3 OR 7.31 [4.06–13.36]; C4 2.89 [1.72–4.96], Firth) and hospital-acquired LOS (C3 OR 9.05 [5.09–16.45]; C4 3.02 [1.78–5.24]); death was not significantly associated. In PIC, by contrast, the higher composite risk was driven primarily by death (C1 OR 2.75 [1.82–4.13]; C3 OR 4.91 [2.49–9.23]; C4 OR 2.62 [1.59–4.23]), whereas NEC and LOS were not associated with trajectory class (Table 3). Thus, the components driving the composite signal differed between databases.

Table 3
Component-wise associations between trajectory class and outcomes (Firth penalized logistic regression)

Outcome	C1 OR (95% CI)	P	C3 OR (95% CI)	P	C4 OR (95% CI)	P
MIMIC-III						
Composite (death/NEC/LOS)	0.90 (0.24– 2.40)	0.845	7.37 (4.79– 11.41)	< 0.001	2.64 (1.80– 3.88)	< 0.001
Death	3.60 (0.79– 13.24)	0.092	1.69 (0.58– 5.07)	0.337	0.75 (0.23– 2.31)	0.622
NEC	0.27 (0.00– 2.06)	0.274	7.31 (4.06– 13.36)	< 0.001	2.89 (1.72– 4.96)	< 0.001
Hospital-acquired LOS	0.29 (0.00– 2.19)	0.301	9.05 (5.09– 16.45)	< 0.001	3.02 (1.78– 5.24)	< 0.001
PIC						
Composite (death/NEC/LOS)	1.50 (1.09– 2.06)	0.014	2.57 (1.41– 4.51)	0.003	1.90 (1.30– 2.72)	< 0.001
Death	2.75 (1.82– 4.13)	< 0.001	4.91 (2.49– 9.23)	< 0.001	2.62 (1.59– 4.23)	< 0.001
NEC	0.53 (0.16– 1.35)	0.196	0.83 (0.09– 3.34)	0.820	1.02 (0.33– 2.60)	0.965
Hospital-acquired LOS	1.01 (0.64– 1.55)	0.965	1.19 (0.43– 2.72)	0.707	1.29 (0.76– 2.09)	0.332
OR, odds ratio; CI, confidence interval. Cell values are OR (95% CI) and P relative to the typical class (C2). Firth penalized logistic regression was used to address sparse-data separation, particularly for rare components in small classes.						

Secondary outcomes

Prolonged hospitalization (> 28 days) was strongly associated with trajectory class in MIMIC-III (C1 OR 0.38 [0.18–0.83]; C3 OR 12.16 [8.44–17.51]; C4 OR 4.01 [3.25–4.93]) but not in PIC (all $p > 0.7$). The protective association of C1 with prolonged hospitalization is consistent with its composition: a quarter of C1 infants were admitted at 72 h–7 d of age and its median admission weight was 2.6 kg, indicating comparatively mature infants admitted for late jaundice with short stays, in contrast to the very preterm infants dominating C3. In PIC, the acute-kidney-injury proxy was associated with C1 (OR 1.41 [1.07–1.87]) and C3 (OR 2.40 [1.40–4.10]) but not C4.

Incremental value of the trajectory phenotype over single-point measures

In the primary analysis population, the trajectory model (A) achieved apparent AUCs of 0.827 (MIMIC-III) and 0.604 (PIC); after bootstrap optimism correction, corrected AUCs were 0.820 and 0.579 (Table S2). In MIMIC-III, on apparent AUCs, model A significantly outperformed the Bhutani risk-zone model (B3; apparent AUC 0.827 vs 0.776, DeLong $p < 0.001$; corrected AUC 0.820 vs 0.768; continuous NRI 0.576 [0.459–0.687]; IDI 0.042 [0.031–0.061]) and showed significant NRI/IDI improvement over the first-value model (B1; NRI 0.198 [0.074–0.337]; IDI 0.017 [0.004–0.038]) without a significant AUC difference (DeLong $p = 0.133$); it did not outperform the peak-value model (B2; apparent 0.827 vs 0.827, DeLong $p = 0.987$; corrected 0.820 vs 0.822; NRI – 0.120); the negative NRI against the peak-value model further indicates no incremental value over the peak in MIMIC-III. In PIC, the AUC difference was significant only versus the peak value (DeLong $p = 0.010$) and borderline versus the first-value and Bhutani models ($p = 0.054$ and 0.051), while continuous NRI was significant against all three (vs B1 0.207 [0.100–0.318]; vs B2 0.147 [0.026–0.253]; vs B3 0.161 [0.056–0.267]) (Fig. 4, Table S3). Restricting to neonates admitted within 72 h or 24 h of birth (S8) preserved these findings: in PIC, the trajectory model improved over first and peak values (DeLong $p = 0.013$ – 0.016) with larger NRI (0.27–0.28), while the absolute AUC remained approximately 0.60.

Sensitivity analyses

Pre-specified sensitivity analyses were broadly consistent with the primary results (Table S4). Refitting the trajectory model on the ≥ 2 -measurement population (S2; $n = 3,644$ MIMIC-III and 2,078 PIC; one infant whose measurements were all physiologically impossible values ≤ 0 was excluded, so 2,078 rather than 2,079 entered the refit) again identified classes with strong associations with the composite outcome (MIMIC-III OR up to 7.16 [4.75–10.80]; PIC OR 1.73–2.35, all significant), indicating that the primary results are not an artefact of the ≥ 3 -measurement threshold; because the refit uses its own class numbering, these results are shown separately in Table S4b. The associations of C3 and C4 with the composite outcome were robust to excluding early-onset sepsis (S5), community-acquired bacteremia (S6), and coagulase-negative staphylococci (S7), and to restricting to early admissions (S8).

In MIMIC-III, all composite events occurred among infants admitted within the first 24 h of life, so S6 and S8 removed only event-free infants and, under quasi-complete separation, the class estimates were identical to the primary analysis to the precision shown in Table S4; in PIC, where events occurred across admission strata, these sensitivities did change the estimates. In MIMIC-III, the C3 association persisted after excluding phototherapy (S9; OR 10.10 [3.30–30.88]), but the C4 association lost significance in this subset (OR 2.23 [0.73–6.82], $p = 0.159$); S9 retained only 1,026 MIMIC-III infants (71% had received phototherapy), and the reduced event count markedly lowered statistical power for the less common C4 class. Posterior-probability-weighted class assignment (S11) confirmed the C3 and C4 associations in both databases. In a minimal-adjustment model including only sex and age at admission (S14), associations were stronger than in the primary analysis in MIMIC-III (C3 OR 15.59 [10.57–22.99]; C4 OR 3.47 [2.40–5.01]) and similar in PIC (C3 OR 2.59 [1.45–4.60]; C4 OR 1.92 [1.33–2.77]) (Table S4), indicating that adjustment for clinical covariates attenuated the estimates, particularly for the immature C3 class. In contrast, adjusting MIMIC-III models for birth weight in the complete-case subset attenuated the C3 and C4 associations to non-significance (S14; C3 OR 1.23 [0.71–2.13]; C4 OR 0.98 [0.63–1.51]), indicating that in MIMIC-III a substantial part of the trajectory-outcome association is attributable to birth weight, because the high-risk classes comprised predominantly very-low-birth-weight infants. In contrast, adding a measure of respiratory support intensity (maximal FiO₂ in the first 24 h; S19, $n = 1,718$) did not eliminate the associations (C3 OR 5.64 [3.51–9.08]; C4 OR 2.17 [1.42–3.31]), indicating that adjustment for a severity proxy other than birth weight did not account for the trajectory-outcome association. E-value analysis (S15) indicated that unmeasured confounding would need to be strong to fully explain the C3 associations (MIMIC-III point/lower-CI E-values 14.39/9.11; PIC 4.54/2.20) but only modest to explain the PIC C1 association (2.37/1.40) (Table S5). The high MIMIC-III C3 E-value should be interpreted in light of the birth-weight sensitivity: because the E-value is computed from the primary-model estimate, in which immaturity was captured only by incompletely coded ICD variables, it quantifies the strength required of confounding beyond that already demonstrated for birth weight, rather than implying robustness to immaturity confounding; S14 shows that a single measured covariate was in fact sufficient to attenuate the association. Excluding outcomes occurring within postnatal days 0–7 (S17) attenuated the PIC C3 association to non-significance (OR 1.77 [0.83–3.75]) while the PIC C4 association remained significant (OR 1.65 [1.08–2.54]); MIMIC-III C3 and C4 associations persisted (S17). In the maturity-proxy subgroup approximating the ≥ 35 -week population in which the Bhutani nomogram is validated (S35), associations in MIMIC-III (birth weight ≥ 2.0 kg; 16 composite events among 1,283 infants) could not be reliably estimated and are reported as not estimable; in PIC (no preterm ICD code; $n = 1,309$), all three non-reference classes remained significantly associated with the composite outcome (C1 OR 2.02 [1.39–2.92]; C3 OR 2.65 [1.42–4.95]; C4 OR 1.88 [1.20–2.94]), and the trajectory model showed borderline-better discrimination than the Bhutani model (AUC 0.637 vs 0.600; DeLong $p = 0.064$) (Table S10). Per-database trajectory refitting confirmed the high-risk persistently low phenotype in MIMIC-III (OR 5.66 [3.72–8.61]) but selected only two classes in PIC, supporting an early-peak high-risk class (OR 2.13 [1.40–3.24]); agreement between pooled and per-database class assignments was moderate (adjusted Rand index 0.23–0.25; Table S6). The Bhutani-zone-from-peak sensitivity (S10) was unstable (complete separation in MIMIC-III; no consistent gradient in PIC) and was

not used for inference. Exploratory unsupervised clustering showed low agreement with the latent classes (ARI approximately 0) and weaker outcome discrimination than the latent-class model, indicating that the class structure is model-dependent and not recoverable by generic clustering (Table S7). Calibration of the trajectory model was acceptable in MIMIC-III (Brier score 0.0581 vs null 0.0664; Brier skill 12.5%; bootstrap-optimism-corrected calibration intercept – 0.11 and slope 0.95) but showed greater over-optimism in PIC (Brier score 0.1392 vs null 0.1433; Brier skill 2.9%; corrected intercept – 0.37 and slope 0.76), consistent with the lower discrimination in that database (Table S8); in MIMIC-III the peak-value model showed marginally better calibration than the trajectory model (Brier skill 14.1% vs 12.5%), mirroring the AUC comparison. In decision-curve analysis, the trajectory model provided higher net benefit than the Bhutani model across all examined thresholds in MIMIC-III, whereas in PIC differences were small and threshold-dependent; net-benefit differences between the trajectory model and the peak-value model were small in both databases (Fig. 5, Table S9).

Summary of robust findings

Across both databases, the most reproducible findings were: (i) a rapid-rise early-peak trajectory (C4) associated with higher composite risk, robust to temporal sensitivity in PIC and to most sensitivities in MIMIC-III, with two explicit exceptions in MIMIC-III: attenuation to non-significance after birth-weight adjustment in the complete-case subset (S14) and after excluding phototherapy (S9, in which the analysis retained only 1,026 infants and was underpowered); and (ii) incremental discrimination of the trajectory phenotype over the Bhutani single-point approach, which must be interpreted with the caveat that the nomogram was applied largely outside its validated ≥ 35 -week range. The persistently low phenotype (C3) was strongly associated with adverse outcomes in both databases in the primary analysis despite the lowest bilirubin values, but was substantially explained by birth weight in MIMIC-III and attenuated after excluding early outcomes in PIC; it is therefore interpreted cautiously as a severity/acuity marker rather than an independent bilirubin-specific effect.

Discussion

Principal findings

In two independent neonatal intensive care databases (MIMIC-III and PIC), latent-class trajectories of total bilirubin over the first postnatal week identified a rapid-rise early-peak phenotype (C4) that was consistently associated with a higher composite outcome (death, NEC, or hospital-acquired LOS) after extensive adjustment and across nearly all pre-specified sensitivity analyses, and a persistently low phenotype (C3) that was strongly associated with adverse outcomes in the primary analysis—despite the lowest bilirubin values of all classes—but whose interpretation was substantially qualified by birth-weight confounding (MIMIC-III) and by attenuation after excluding early outcomes (PIC). The trajectory phenotype provided incremental discrimination over the classic hour-specific Bhutani single-point

approach, applied here largely outside its validated ≥ 35 -week range, and over the first value; in MIMIC-III it matched, but did not exceed, the peak value.

Comparison with previous studies

The most directly comparable study is a single-center MIMIC-III analysis in which a higher TSB at NICU admission was associated with lower odds of acute kidney injury, illustrating that single-point bilirubin measures carry prognostic information beyond jaundice [10]. Our results extend this literature in three ways. First, by characterizing the temporal shape of early bilirubin rather than a single admission value, we identified a rapid-rise early-peak phenotype that was associated with the composite outcome in the same direction in both databases; whether single-point measures show comparable cross-database consistency was not formally tested. Second, our findings are consistent with the dual biology of bilirubin—antioxidant protection at low concentrations [4–6] and neurotoxicity or cholestasis-associated severity at high or poorly cleared concentrations [7, 8, 9]—and with the recognition that sepsis-related cholestasis marks illness severity in neonates [8, 9]. Third, trajectory approaches are gaining acceptance for neonatal biomarkers [13, 14]; our results add bilirubin to this body of work while underscoring the importance of compositional scrutiny, because the phenotype-outcome associations were partly confounded by immaturity. The persistently low phenotype observed in the most immature infants parallels the low bilirubin values described in extremely preterm cohorts and is best interpreted as a marker of severity rather than a benign pattern [5, 15]. Finally, our analysis of single-point comparators echoes earlier work comparing hour-specific risk-assessment approaches in healthy newborns [16] and highlights that the choice of comparator determines the apparent incremental value of trajectory modeling.

The rapid-rise early-peak phenotype as the most robust finding

The C4 class was the most reproducible high-risk phenotype: it remained significant in PIC after excluding outcomes within the exposure window (S17), was not explained by birth weight in PIC, and persisted across most sensitivities in MIMIC-III. This phenotype is clinically plausible: a rapid early rise in bilirubin may reflect hemolysis, sepsis-related cholestasis, or a sick infant in whom the bilirubin trajectory tracks evolving systemic illness. From a clinical standpoint, identifying infants whose bilirubin is rising quickly and peaking early could prompt earlier reassessment of the underlying cause (e.g., hemolysis or infection) rather than treatment of the bilirubin value alone.

The persistently low phenotype as a severity marker

The C3 class had the highest composite-outcome risk in the primary analysis (MIMIC-III OR 7.45; PIC OR 2.55) despite having the lowest first and peak TSB values of all classes in both databases. In MIMIC-III, this class comprised predominantly very premature, very-low-birth-weight infants (97.7% preterm; median admission weight 0.8 kg; median six TSB measurements), and adjusting for birth weight attenuated the association to non-significance, indicating that in that database much of the signal is attributable to extreme immaturity and low birth weight rather than to the bilirubin trajectory per se. Low

bilirubin in such infants is physiologically plausible—reduced hepatic bilirubin production and uptake, enhanced phototherapy, and short survival windows truncating the postnatal curve—and parallels reports in extremely preterm cohorts [5, 15]. Low bilirubin may further indicate reduced antioxidant reserve in the most immature infants: a prospective study found higher bilirubinemia to be associated with less severe retinopathy of prematurity, an oxidative-stress-driven morbidity, consistent with a protective role of bilirubin in this population [17]. In PIC, the C3 class was compositionally different (mostly term, normal-weight infants; 9.8% preterm) yet also had high mortality (23.0%); its association attenuated after excluding outcomes occurring within the exposure window, suggesting that some of the association reflects concurrent, early-onset severe illness. We therefore interpret C3 as a marker of acuity and severity—“low bilirubin in the sickest infants”—rather than an independent bilirubin-specific effect. These findings underscore that trajectory phenotypes can aggregate heterogeneous populations across centers and that compositional differences must be examined before attributing phenotype-outcome associations to the biomarker itself.

Incremental value over single-point measurement

The trajectory phenotype improved discrimination over the classic Bhutani risk zone in MIMIC-III (corrected AUC 0.820 vs 0.768; continuous NRI 0.58) and, on continuous NRI, in PIC (NRI 0.16), supporting the notion that temporal information adds risk-stratification value beyond an hour-specific single value [1, 2, 10]. This comparison must be interpreted with caution, however, because the Bhutani nomogram was developed and validated in healthy infants born at ≥ 35 weeks [1, 2], whereas 78–98% of the MIMIC-III cohort were preterm and the nomogram was therefore applied largely outside its validated range; the maturity-proxy subgroup analyses (S35; Table S10) were limited by sparse events, and gestational age was unavailable in both final exports, so the validated population could not be examined directly. In MIMIC-III the trajectory model did not outperform the peak value (corrected AUC 0.820 vs 0.822), suggesting that, in a population with intensive serial monitoring, the peak may already capture much of the trajectory information relevant to the composite outcome. In PIC, where monitoring and admission patterns differ, the trajectory added more clearly over single points (significant continuous NRI against all three). Net-benefit (decision-curve) analyses favored the trajectory model over the Bhutani model across all examined thresholds in MIMIC-III and were threshold-dependent in PIC; net-benefit differences between the trajectory model and the single-point models were small and did not consistently favour the trajectory model (e.g., in MIMIC-III the peak-value model had marginally higher net benefit at thresholds of 0.10–0.20), consistent with the AUC results (Fig. 5, Table S9). These results qualify, rather than negate, the incremental-value hypothesis and indicate that the added value of trajectory phenotyping may depend on monitoring intensity, population, and the choice of comparator.

Component drivers and cross-database differences

The components driving the composite signal differed between databases: NEC and hospital-acquired LOS in MIMIC-III versus death in PIC. This difference likely reflects different case-mix and care patterns (e.g., a referral center with high neonatal mortality in PIC) and is consistent with the marked difference in overall discrimination between databases (MIMIC-III AUC 0.82 vs PIC 0.60). The delayed high-peak class

(C1) showed inconsistent direction across databases (no association in MIMIC-III; higher risk in PIC), further supporting that the pooled four-class structure should be interpreted as a shared analytic framework rather than as identical phenotypes in each center; per-database refitting confirmed only two classes in PIC. These caveats are inherent to two-database trajectory work and are reported transparently.

Clinical implications

If confirmed prospectively, trajectory-based risk stratification could identify neonates at higher risk of death, NEC, or late-onset sepsis earlier than single-point assessment, and could support phenotype-directed monitoring (e.g., rapid-early-rise cases warranting evaluation for hemolysis and infection). Because the composite outcome is multifactorial, trajectory phenotypes should be viewed as stratification tools for closer surveillance rather than as causal or treatable targets; in particular, the persistently low phenotype should be recognized as a severity marker rather than a reassuring pattern. Finally, the observation that in MIMIC-III all composite events occurred among infants admitted within the first 24 h of life indicates that the outcome burden in that database was concentrated in in-born, predominantly preterm infants, whereas later-admitted infants (predominantly C1, admitted for late jaundice) had essentially no events; this population structure underscores that trajectory phenotypes aggregate different clinical populations across centers and that case-mix must be considered when comparing phenotype-outcome associations.

Strengths and limitations

Strengths include two independent databases from different healthcare systems, prospective preregistration, reproducible analysis scripts, a comprehensive pre-specified sensitivity strategy (including the ≥ 2 -measurement threshold, Firth penalization, bootstrap optimism correction, temporal and early-admission sensitivities, posterior-probability weighting, E-value analysis [18], and per-database refitting), and supplementary calibration and decision-curve analyses. E-values were interpreted as quantifying the strength of unmeasured confounding beyond that captured by measured covariates such as birth weight, in keeping with the birth-weight sensitivity result.

Limitations include the retrospective design; the pooled trajectory model (with per-database refitting as a sensitivity); unavailable gestational age and Apgar score in the final MIMIC-III export, which precluded a direct evaluation of the Bhutani-validated ≥ 35 -week population (addressed by symmetric covariates, a birth-weight sensitivity, a maturity-proxy subgroup, and E-value analysis); unavailable gestational age and phototherapy coding in PIC; ICD-based preterm/low-birth-weight covariates that are incompletely captured (e.g., only 16.1% of the MIMIC-III C3 class had a low-birth-weight ICD code despite a median admission weight of 0.8 kg), which probably attenuates adjustment for immaturity in the primary model and contributes to the birth-weight attenuation observed in S14; NEC without a diagnostic date (temporal sensitivity approximated using death/LOS times); incomplete prescription coverage; the acute-kidney-injury proxy; moderate absolute discrimination, particularly in PIC; cross-database differences in class composition and the modest agreement between pooled and per-database class assignments (adjusted

Rand index 0.23–0.25); the exploratory clustering analysis showing low agreement with the latent classes (ARI approximately 0): while the latent classes discriminated outcomes better than the generic clustering solutions, the low agreement also indicates that the class structure is model-dependent and not recoverable by standard clustering; and differences in TSB measurement methods and analytical platforms across databases (with units converted by dividing by 17.1), which cannot be fully excluded as a source of residual heterogeneity. TSB measurement frequency also differed systematically across classes and reflects both illness severity and survival duration, so class assignment is partly a function of monitoring intensity (an informative observation process [19]); the ≥ 2 -measurement threshold analysis and measurement-count comparisons were performed to address this, but residual informative observation cannot be excluded. The remaining preregistered analyses not performed (0-3-day window, > 14-day prolonged stay, multiple imputation, drug covariates, exposure-window truncation, and time-to-event analyses) are acknowledged in the Methods.

Conclusions

Latent-class trajectories of early neonatal bilirubin identified a rapid-rise early-peak phenotype associated with higher composite adverse outcomes in two independent databases, with a trajectory model that provided incremental discrimination over the classic hour-specific Bhutani risk zone—applied here largely outside its validated range—and over the first value, while matching, but not exceeding, the peak value in MIMIC-III. The trajectory-outcome association was partly explained by birth weight and by acute illness in some subgroups, underscoring the need for compositional and temporal scrutiny when interpreting biomarker trajectories. Prospective validation, ideally in a ≥ 35 -week population, is warranted before trajectory-based phenotyping can inform clinical risk stratification.

Methods

Study design, setting, and preregistration

This was a retrospective, two-center cohort study using data from the Medical Information Mart for Intensive Care III (MIMIC-III) database (Beth Israel Deaconess Medical Center, Boston, MA, USA; admissions 2001–2008 in this extract) and the Pediatric Intensive Care (PIC) database, a publicly available, pediatric-specific intensive care database established by the Children's Hospital, Zhejiang University School of Medicine (Hangzhou, China; <http://pic.nbscn.org>) and structured in accordance with the MIMIC data schema (admissions 2010–2018 in this extract). Both databases contain admission-level demographic, diagnostic, prescription, laboratory, and vital-sign data for critically ill neonates. The analysis plan, including exposures, outcomes, covariates, and the full sensitivity strategy, was preregistered on the Open Science Framework (OSF; <https://osf.io/x7t42>) before the outcome analyses were conducted. The study is reported in accordance with the STROBE [20] and RECORD [21] statements and in line with recent reviews of trajectory-analysis methods and reporting [13]. MIMIC-III use was governed by its credentialed data-use agreement (PhysioNet). The PIC database was established with approval from the Institutional Review Board of the Children's Hospital, Zhejiang University School of

Medicine (approval no. 2019_IRB_052), and the requirement for individual informed consent was waived because the study did not affect clinical care and all protected health information was de-identified. Because this study used only publicly available, de-identified data, no additional ethics approval was required. All procedures were performed in accordance with the Declaration of Helsinki.

Study population

We included neonates admitted at a postnatal age < 28 days. Because the exposure is the bilirubin trajectory during the first postnatal week, the primary analysis population comprised neonates with at least three TSB measurements on postnatal days 0–7 (MIMIC-III, $n = 2,951$; PIC, $n = 1,817$). The full cohort comprised 8,059 (MIMIC-III) and 2,953 (PIC) neonates admitted before 28 days of age. Neonates without any TSB measurement on days 0–7 (e.g., those admitted after the first postnatal week) were excluded from the trajectory analysis but remain in the full-cohort description (Fig. 1). By construction, the trajectory population was therefore restricted to infants hospitalized during the first postnatal week, including those admitted within the first 24 h (MIMIC-III 2,842; PIC 1,129) and those admitted up to day 7.

Exposure: latent-class trajectories of bilirubin

TSB measurements (MIMIC-III item 50885, mg/dL; PIC items 5075/5255, $\mu\text{mol/L}$ converted to mg/dL by dividing by 17.1) were aligned to a postnatal-age axis (hours after birth). Latent-class trajectory models were fitted to log-transformed TSB over postnatal hours 0–168 using a latent-class linear mixed model (R package *lcmm* [12]; fixed and class-varying linear and quadratic time terms, with a random intercept; single processor for reproducibility). Models with two to five classes were estimated. The number of classes was selected by the pre-specified rule of best BIC combined with a minimum class proportion $\geq 5\%$ and a mean maximum posterior probability ≥ 0.7 . The BIC-preferred five-class model was rejected because two classes each comprised < 5% of the population; the four-class model satisfied all pre-specified criteria and was retained. The mean maximum posterior probability of the four-class model was 0.773, only modestly above the pre-specified threshold of 0.7, indicating non-trivial classification uncertainty for some infants; we address this through posterior-probability-weighted assignment in a sensitivity analysis (S11). Class membership was assigned by the modal posterior probability. To provide a shared phenotype framework for cross-database comparison, the trajectory model was fitted on the combined two-database population; the model was additionally refitted separately within each database to assess robustness to this modeling choice (supplementary material).

Single-point comparators

Within the same postnatal days 0–7 window, we derived three single-point measures from the same TSB values used for the trajectories: (i) the first value; (ii) the peak value; and (iii) the Bhutani hour-specific risk zone, computed from the first value and its postnatal hour using the hour-specific percentile nomogram (40th, 75th, and 95th percentiles; digitized reference coordinates at 12–148 h with linear interpolation; values beyond 148 h and infants born before 35 weeks of gestation were considered extrapolation, because the nomogram was developed and validated in healthy infants born at ≥ 35 weeks

[1, 2] and its use in more preterm infants is not recommended [15]). We retained the hour-specific risk zone as the single-point comparator for three reasons: it is mechanically computable from the first value and its postnatal hour; it does not require gestational age, which was unavailable in both final exports; and it mirrors the definitions used in prior single-point analyses of bilirubin and neonatal outcomes [10]. Although the 2022 American Academy of Pediatrics guideline [3] has moved management toward treatment-threshold curves, the risk-zone framework remains the reference predischarge risk-assessment tool and is the appropriate comparator for trajectory phenotyping.

Outcome

The primary outcome was the composite of in-hospital death, NEC, or hospital-acquired LOS, mirroring composite outcomes used in neonatal trials and core outcome sets [22–24]. Bilirubin exposure timing was expressed on a postnatal-age axis, whereas sepsis timing used admission-relative anchors ($t_0 + 48$ h, where t_0 is the time of ICU or hospital admission, whichever came first), consistent with hospital-acquired-infection definitions [25–27]; these two time axes are intentionally distinct. NEC was ascertained from ICD-9 777.x (MIMIC-III) or ICD-10 P77 or the diagnosis text “necrotizing enterocolitis” (PIC). Hospital-acquired LOS was defined as a positive blood culture collected at a postnatal age ≥ 72 h and at ≥ 48 h after t_0 ; early-onset sepsis (positive culture at age < 72 h) and community-acquired/present-on-admission bacteremia (positive culture at age ≥ 72 h but < 48 h after t_0) were treated as covariates rather than outcome events [25–27]. The three components were also analyzed separately. Secondary outcomes included prolonged hospitalization (> 28 days) and, in PIC only, an acute-kidney-injury proxy (≥ 2 serum creatinine measurements on postnatal days 0–7). Infants without any blood culture were coded as not having culture-confirmed sepsis.

Covariates

The primary models adjusted for sex; postnatal age at admission (continuous and categorical: < 24 h, $24–72$ h, 72 h– 7 d, > 7 d); ICD-based preterm birth and low birth weight; congenital anomalies [28]; early-onset sepsis; community-acquired/present-on-admission bacteremia; and vasoactive drug use within the first 24 h. The same covariate set was used in both databases (symmetric adjustment). Because gestational age and Apgar score were unavailable in the final MIMIC-III export, they were not included in the primary model; birth weight and maximal FiO_2 in the first 24 h were examined in pre-specified sensitivity analyses (S14, S19). Admission weight, although more completely recorded, was not used in the primary model because it reflects weight after transfer and any pre-admission interventions rather than birth weight per se, and because the primary covariate set was fixed at preregistration; birth weight was therefore examined as a pre-specified sensitivity analysis.

Statistical analysis

For each database, logistic regression models related trajectory class to the composite outcome, with the largest (reference) class as the comparator, adjusted for the covariates described in the Covariates subsection. To address quasi-complete separation, we additionally fitted Firth penalized-logistic models.

For the core incremental analysis, in the primary analysis population with available single-point values, we compared discrimination of the trajectory model (model A: trajectory class plus covariates) with three single-point models that contained the identical covariate set—B1 (first TSB value plus covariates), B2 (peak TSB value plus covariates), and B3 (Bhutani risk zone plus covariates)—so that any difference in discrimination is attributable to the exposure representation alone. Comparisons used the area under the receiver-operating-characteristic curve (AUC) with DeLong tests, and the continuous net reclassification improvement (NRI) and integrated discrimination improvement (IDI) with bootstrap 95% confidence intervals [29]. Apparent AUCs were corrected for optimism by internal bootstrap validation (200 resamples); DeLong tests were performed on apparent AUCs, and corrected AUCs are reported alongside for transparency. In addition, we evaluated calibration (Brier score; calibration intercept and slope from logistic recalibration) [30, 31] and decision-curve analysis (net benefit across threshold probabilities) [32]. Because continuous NRI can be inflated by small risk shifts that are not clinically meaningful, NRI was interpreted as secondary to AUC, calibration, and net-benefit comparisons, and its limitations are acknowledged [33]. Since the primary analysis tested one pre-specified hypothesis per database (higher composite risk in the rapid-rise early-peak class), the additional class, outcome, and sensitivity comparisons were considered supportive and P values were not adjusted for multiple testing.

Pre-specified sensitivity analyses included: the ≥ 2 -measurement trajectory threshold (S2, requiring refitting of the trajectory model); excluding early-onset sepsis (S5); excluding community-acquired bacteremia (S6); excluding coagulase-negative staphylococci from the LOS definition (S7); restricting to early admissions (< 72 h and < 24 h; S8); excluding phototherapy (MIMIC-III only) and a phototherapy-by-class interaction (S9); the Bhutani zone derived from the peak value (S10); posterior-probability-weighted class assignment (S11); a minimal-adjustment model including only sex and age at admission, with a MIMIC-III enhancement adding birth weight (S14); E-value analysis for unmeasured confounding [18] (S15); excluding outcomes occurring within postnatal days 0–7 to ensure exposure precedes outcome (S17); per-database trajectory refitting; an exploratory unsupervised-clustering validation of the latent classes (S16); and a maturity-proxy subgroup approximating the ≥ 35 -week population in which the Bhutani nomogram is validated (S35; MIMIC-III birth weight ≥ 2.0 kg, PIC absence of a preterm ICD code), because gestational age was unavailable in both final exports.

A small number of preregistered analyses were not performed and are acknowledged as limitations: the 0-3-day window (S3) and exposure-window truncation (S18) require additional trajectory refitting beyond the ≥ 2 -measurement analysis; the > 14 -day prolonged-stay definition (S12) requires a discharge-time variable; multiple imputation (S4) was not applied; drug-covariate adjustment (S13) was deferred to a separate pharmaco-epidemiological analysis; and time-to-event (Cox) analyses were not performed because discharge time was unavailable for censoring. All analyses used R (version 4.4.1) with a fixed random seed for reproducibility; analysis scripts are available on OSF (<https://osf.io/x7t42>).

Abbreviations

AUC

area under the receiver-operating-characteristic curve

CI
confidence interval
DCA
decision-curve analysis
EOS
early-onset sepsis
FiO₂
fraction of inspired oxygen
GBTM
group-based trajectory modeling
ICD
International Classification of Diseases
IDI
integrated discrimination improvement
IQR
interquartile range
LCMM
latent-class linear mixed model
LOS
late-onset sepsis
MIMIC-III
Medical Information Mart for Intensive Care III
NEC
necrotizing enterocolitis
NRI
net reclassification improvement
OR
odds ratio
OSF
Open Science Framework
PIC
Pediatric Intensive Care database
TSB
total serum bilirubin.

Declarations

Data Availability

All data analysed in this study are publicly available secondary databases. MIMIC-III (version 1.4) is available through a credentialed data-use agreement on PhysioNet at <https://physionet.org/content/mimiciii/>. The PIC database is available after registration and approval through the Children's Hospital, Zhejiang University School of Medicine at <http://pic.nbscn.org>. The preregistered analysis plan and all analysis scripts are openly accessible (no login required) on the Open Science Framework at <https://osf.io/x7t42> (registration DOI: <https://doi.org/10.17605/OSF.IO/X7T42>).

Acknowledgements

The authors thank the MIMIC-III project team and the PIC database team for the publicly available datasets.

Author contributions

J.L. conceived and designed the study, developed the methodology, performed the formal analysis, wrote the original draft, and supervised the project. M.W. contributed to methodology and data curation. H.X. contributed to data curation and validation. S.Z. contributed to methodology and prepared the figures. All authors reviewed and edited the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

This secondary analysis used only publicly available, de-identified data and required no additional ethics approval. MIMIC-III was used under its credentialed data-use agreement (PhysioNet); PIC was established with IRB approval of the Children's Hospital, Zhejiang University School of Medicine (approval no. 2019_IRB_052) with waived informed consent. All procedures followed the Declaration of Helsinki.

Consent for publication

Not applicable (no individually identifiable data are presented).

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Use of AI-assisted technologies

AI-assisted tools were used for statistical programming support and language editing. The authors did not use generative AI to generate scientific content, interpretation, or conclusions; they verified all analyses and text and take full responsibility. No AI tool is listed as an author.

Additional information

Correspondence and requests for materials should be addressed to J.L. Supplementary Information accompanies the online version of this paper.

References

1. Bhutani, V. K., Johnson, L. & Sivieri, E. M. Predictive ability of a predischarge hour-specific serum bilirubin for subsequent significant hyperbilirubinemia in healthy term and near-term newborns. *Pediatrics* **103**, 6–14. <https://doi.org/10.1542/peds.103.1.6> (1999).
2. American Academy of Pediatrics Subcommittee on Hyperbilirubinemia. Management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation. *Pediatrics* **114**, 297–316. <https://doi.org/10.1542/peds.114.1.297> (2004).
3. Kemper, A. R. et al. Clinical practice guideline revision: management of hyperbilirubinemia in the newborn infant 35 or more weeks of gestation. *Pediatrics* **150**, e2022058859. <https://doi.org/10.1542/peds.2021-051220> (2022).
4. Hegyi, T., Goldie, E. & Hiatt, M. The protective role of bilirubin in oxygen-radical diseases of the preterm infant. *J. Perinatol.* **14**, 296–300 (1994).
5. Sedlak, T. W. & Snyder, S. H. Bilirubin benefits: cellular protection by a biliverdin reductase antioxidant cycle. *Pediatrics* **113**, 1776–1782. <https://doi.org/10.1542/peds.113.6.1776> (2004).
6. Endesfelder, S., Schmitz, T. & Bühner, C. Bilirubin exerts protective effects on alveolar type II pneumocytes in an in vitro model of oxidative stress. *Int. J. Mol. Sci.* **25**, 5323. <https://doi.org/10.3390/ijms25105323> (2024).
7. Watchko, J. F. & Tiribelli, C. Bilirubin-induced neurologic damage: mechanisms and management approaches. *N Engl. J. Med.* **369**, 2021–2030. <https://doi.org/10.1056/NEJMra1301704> (2013).
8. Chand, N. & Sanyal, A. J. Sepsis-induced cholestasis. *Hepatology* **45**, 230–241. <https://doi.org/10.1002/hep.21480> (2007).
9. Khalil, S., Shah, D., Faridi, M. M. A., Kumar, A. & Mishra, K. Prevalence and outcome of hepatobiliary dysfunction in neonatal septicaemia. *J. Pediatr. Gastroenterol. Nutr.* **54**, 218–222. <https://doi.org/10.1097/MPG.0b013e318233d33d> (2012).
10. Zhou, H. Total bilirubin level is associated with acute kidney injury in neonates admitted to the neonatal intensive care units: based on MIMIC-III database. *Eur. J. Pediatr.* **183**, 4235–4241. <https://doi.org/10.1007/s00431-024-05682-5> (2024).
11. Nagin, D. S. *Group-Based Modeling of Development* (Harvard University Press, 2005). <https://doi.org/10.4159/9780674041318>
12. Proust-Lima, C., Philipps, V. & Liqueur, B. Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. *J. Stat. Softw.* **78**, 1–56. <https://doi.org/10.18637/jss.v078.i02> (2017).

13. Pollington, F. et al. Evaluation of trajectory analysis for disease risk assessment: a scoping review. *J. Am. Med. Inf. Assoc.* **33**, 521–535. <https://doi.org/10.1093/jamia/ocaf208> (2026).
14. Dai, K. et al. Group-based trajectory modelling and interpretable machine learning to identify factors associated with neurodevelopmental trajectories in preterm infants. *BMC Pediatr.* **26**, 115. <https://doi.org/10.1186/s12887-025-06476-w> (2026).
15. Maisels, M. J., Watchko, J. F., Bhutani, V. K. & Stevenson, D. K. An approach to the management of hyperbilirubinemia in the preterm infant less than 35 weeks of gestation. *J. Perinatol.* **32**, 660–664. <https://doi.org/10.1038/jp.2012.71> (2012).
16. Keren, R. et al. Identifying newborns at risk of significant hyperbilirubinaemia: a comparison of two recommended approaches. *Arch. Dis. Child.* **90**, 415–421. <https://doi.org/10.1136/ad.2004.063461> (2005).
17. Rallabandi, S. & Amin, S. B. Bilirubinemia and retinopathy of prematurity in infants ≤ 29 weeks' gestational age. *Pediatr. Res.* **97**, 1065–1071. <https://doi.org/10.1038/s41390-024-03479-8> (2025).
18. VanderWeele, T. J. & Ding, P. Sensitivity analysis in observational research: introducing the E-value. *Ann. Intern. Med.* **167**, 268–274. <https://doi.org/10.7326/M16-2607> (2017).
19. Diggle, P. & Kenward, M. G. Informative drop-out in longitudinal data analysis. *J. R Stat. Soc. Ser. C Appl. Stat.* **43**, 49–73. <https://doi.org/10.2307/2986113> (1994).
20. von Elm, E. et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* **370**, 1453–1457. [https://doi.org/10.1016/S0140-6736\(07\)61602-X](https://doi.org/10.1016/S0140-6736(07)61602-X) (2007).
21. Benchimol, E. I. et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* **12**, e1001885. <https://doi.org/10.1371/journal.pmed.1001885> (2015).
22. Schouten, E. et al. Standardized outcome measures for preterm and hospitalized neonates: an ICHOM standard set. *Neonatology* **119**, 443–454. <https://doi.org/10.1159/000522318> (2022).
23. Webbe, J. W. H. et al. Core outcomes in neonatology: development of a core outcome set for neonatal research. *Arch. Dis. Child. Fetal Neonatal Ed.* **105**, 425–431. <https://doi.org/10.1136/archdischild-2019-317501> (2020).
24. Morowitz, M. J. et al. The NICU Antibiotics and Outcomes (NANO) trial: a randomized multicenter clinical trial assessing empiric antibiotics and clinical outcomes in newborn preterm infants. *Trials* **23**, 428. <https://doi.org/10.1186/s13063-022-06352-3> (2022).
25. Giannoni, E. et al. Neonatal sepsis of early onset, and hospital-acquired and community-acquired late onset: a prospective population-based cohort study. *J. Pediatr.* **201**, 106–114e4. <https://doi.org/10.1016/j.jpeds.2018.05.048> (2018).
26. Centers for Disease Control and Prevention. National Healthcare Safety Network (NHSN) Patient Safety Component Manual. (2026). <https://www.cdc.gov/nhsn/psc-manual/index.html> (accessed 10 August 2026).

27. Zhu, M. L. et al. Comparative analysis of the pathogens responsible for hospital acquired and community acquired late onset neonatal septicemia. *Zhonghua Er Ke Za Zhi*. **46**, 124–127. <https://doi.org/10.3760/cma.j.issn.0578-1310.2008.02.114> (2008).
28. Chughtai, A. A. et al. Association between congenital anomalies and late-onset bacterial infections in neonates admitted to neonatal intensive care units in Australia and New Zealand: a population-based cohort study. *Neonatology* **122**, 95–105. <https://doi.org/10.1159/000540276> (2025).
29. Pencina, M. J., D'Agostino, R. B., Sr., D'Agostino, R. B. Jr. & Vasan, R. S. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* **27**, 157–172. <https://doi.org/10.1002/sim.3010> (2008).
30. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138. <https://doi.org/10.1097/EDE.0b013e3181c30fb2> (2010).
31. Brier, G. W. Verification of forecasts expressed in terms of probability. *Mon Weather Rev.* **78**, 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078%3C0001:VOFFEP%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFFEP%3E2.0.CO;2) (1950).
32. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* **26**, 565–574. <https://doi.org/10.1177/0272989X06295361> (2006).
33. Kerr, K. F. et al. Net reclassification indices for evaluating risk-prediction instruments: a critical review. *Epidemiology* **25**, 114–121. <https://doi.org/10.1097/EDE.000000000000018> (2014).

Figures

Cohort selection flow

Neonates admitted at <28 days of age, MIMIC-III and PIC databases

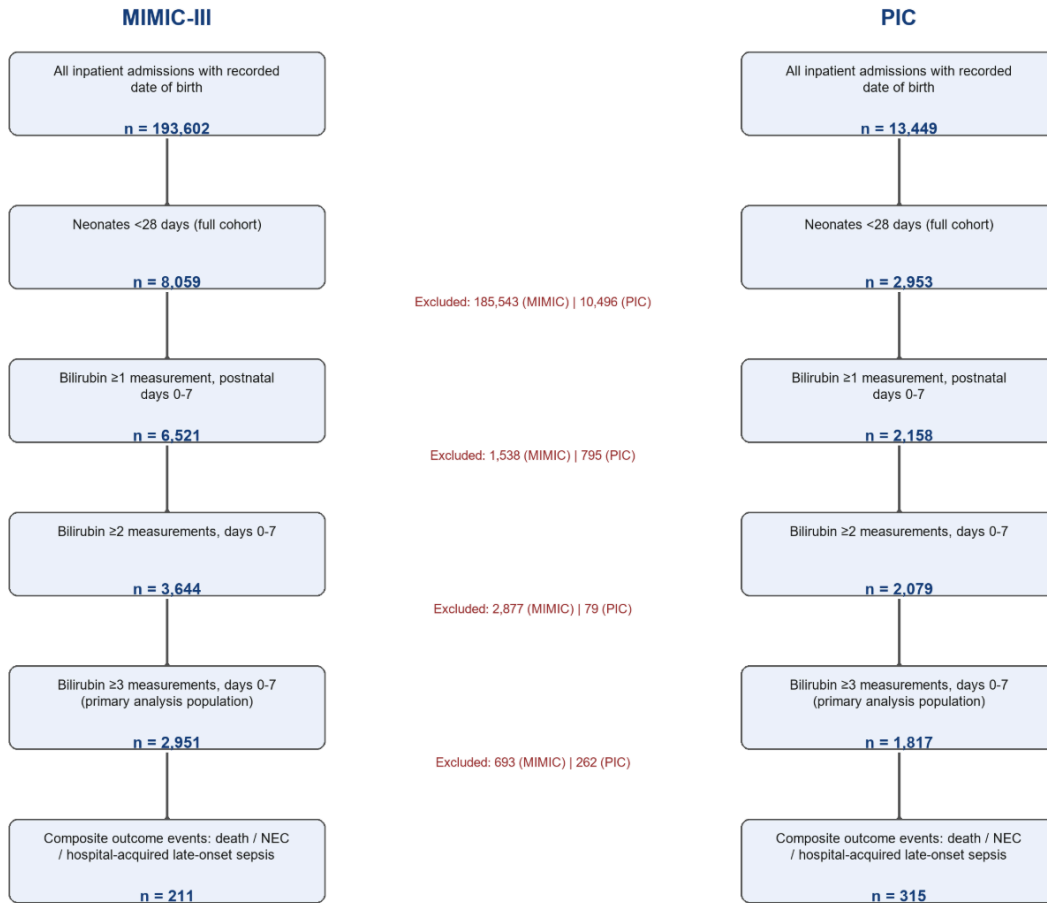


Figure 1

Cohort flow diagram. Full cohort (neonates admitted at <28 days of age) and trajectory analysis population (at least three total bilirubin measurements on postnatal days 0-7) in MIMIC-III and PIC.

Bilirubin trajectory classes (k=4, pooled model, postnatal days 0-7)

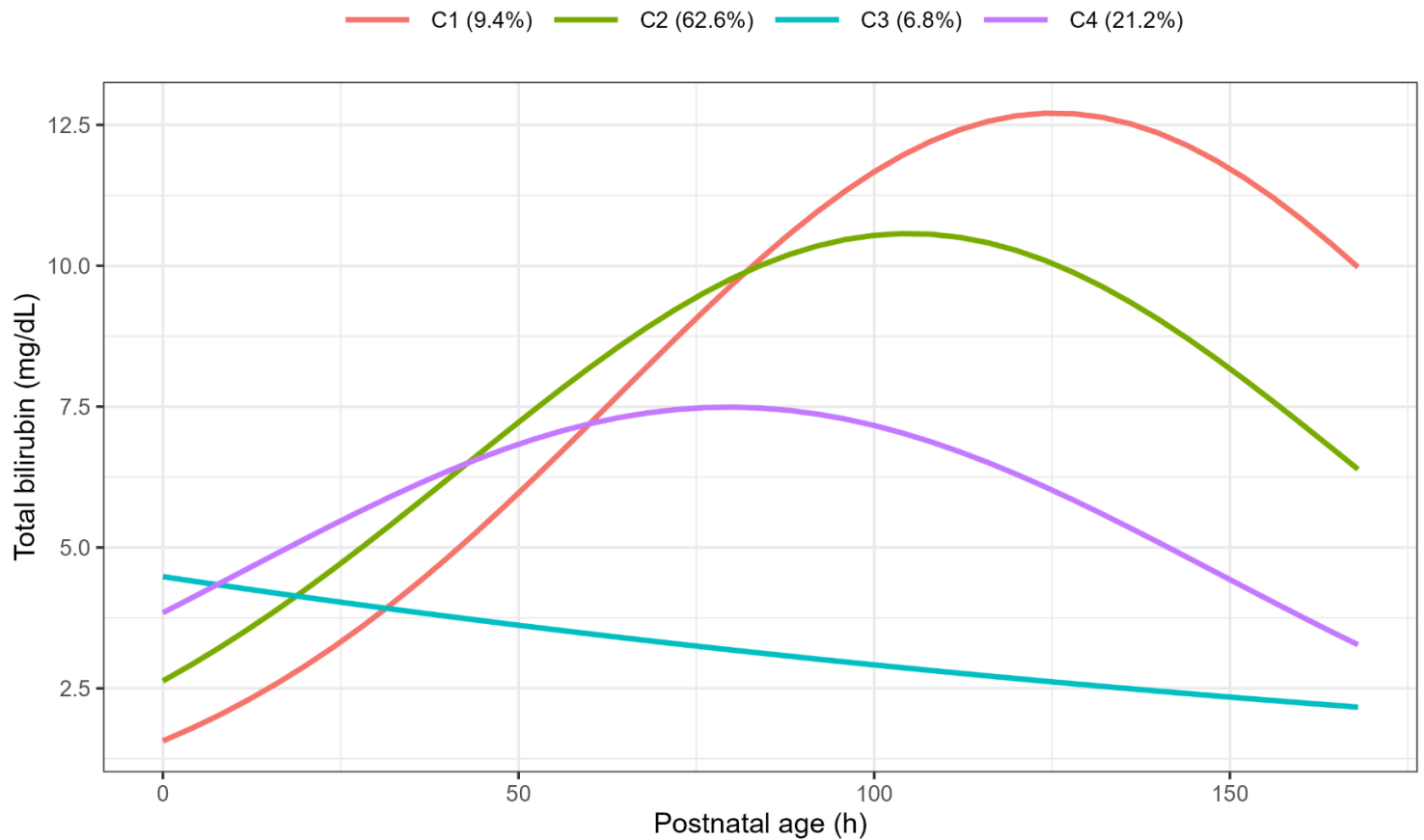


Figure 2

Latent-class trajectories of total bilirubin (mg/dL; back-transformed from the log scale used in modeling) on the postnatal-age axis (hours after birth), four-class solution fitted on the combined two-database population. C1, delayed high peak; C2, typical; C3, persistently low; C4, rapid-rise early peak.

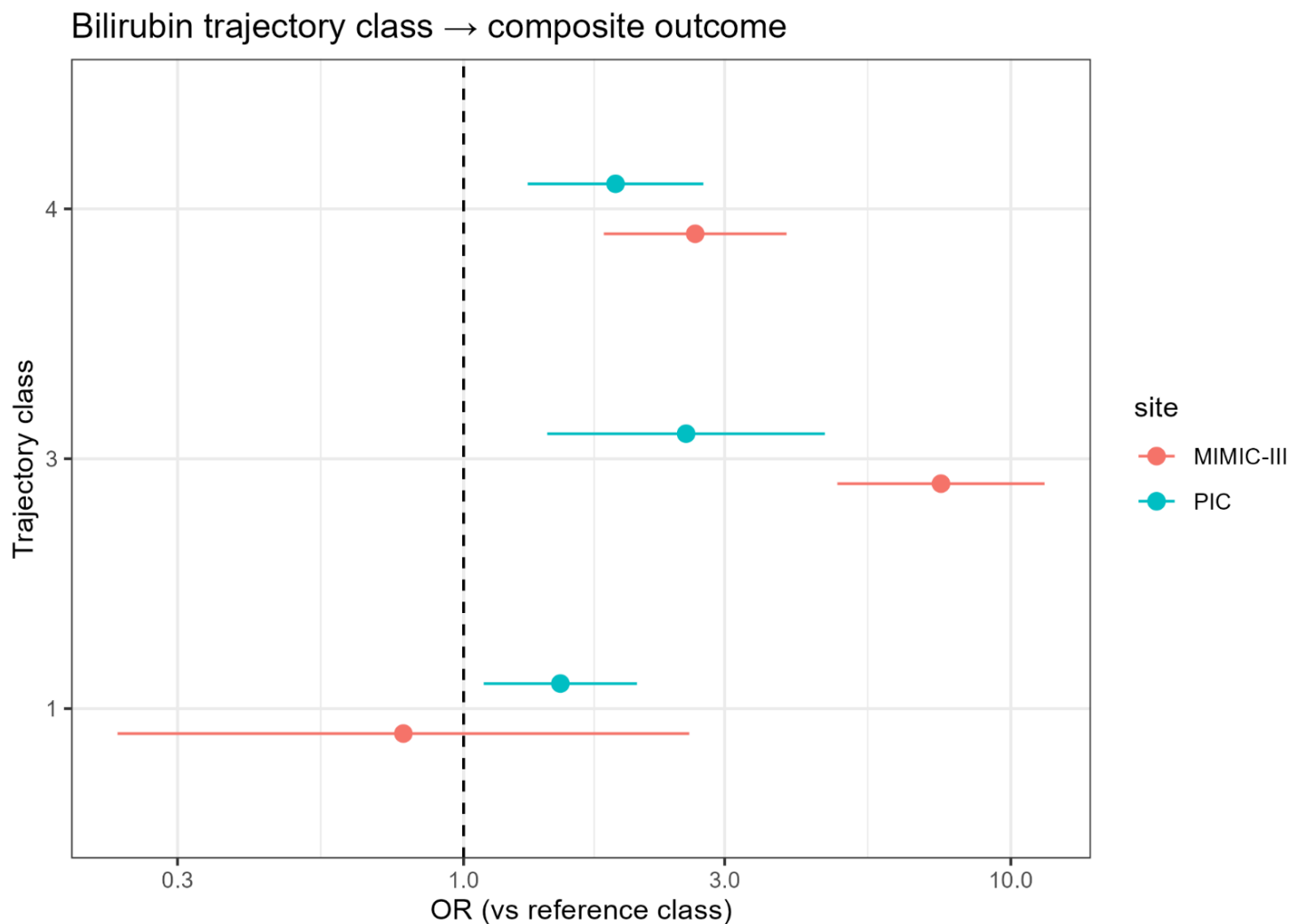


Figure 3

Adjusted odds ratios (95% CI) for the composite outcome (death/NEC/LOS) by trajectory class relative to the typical class (C2), in MIMIC-III and PIC. Logistic regression adjusted for sex, postnatal age at admission, preterm birth/low birth weight, congenital anomalies, early-onset sepsis, community-acquired bacteremia, and vasoactive drug use within 24 h.

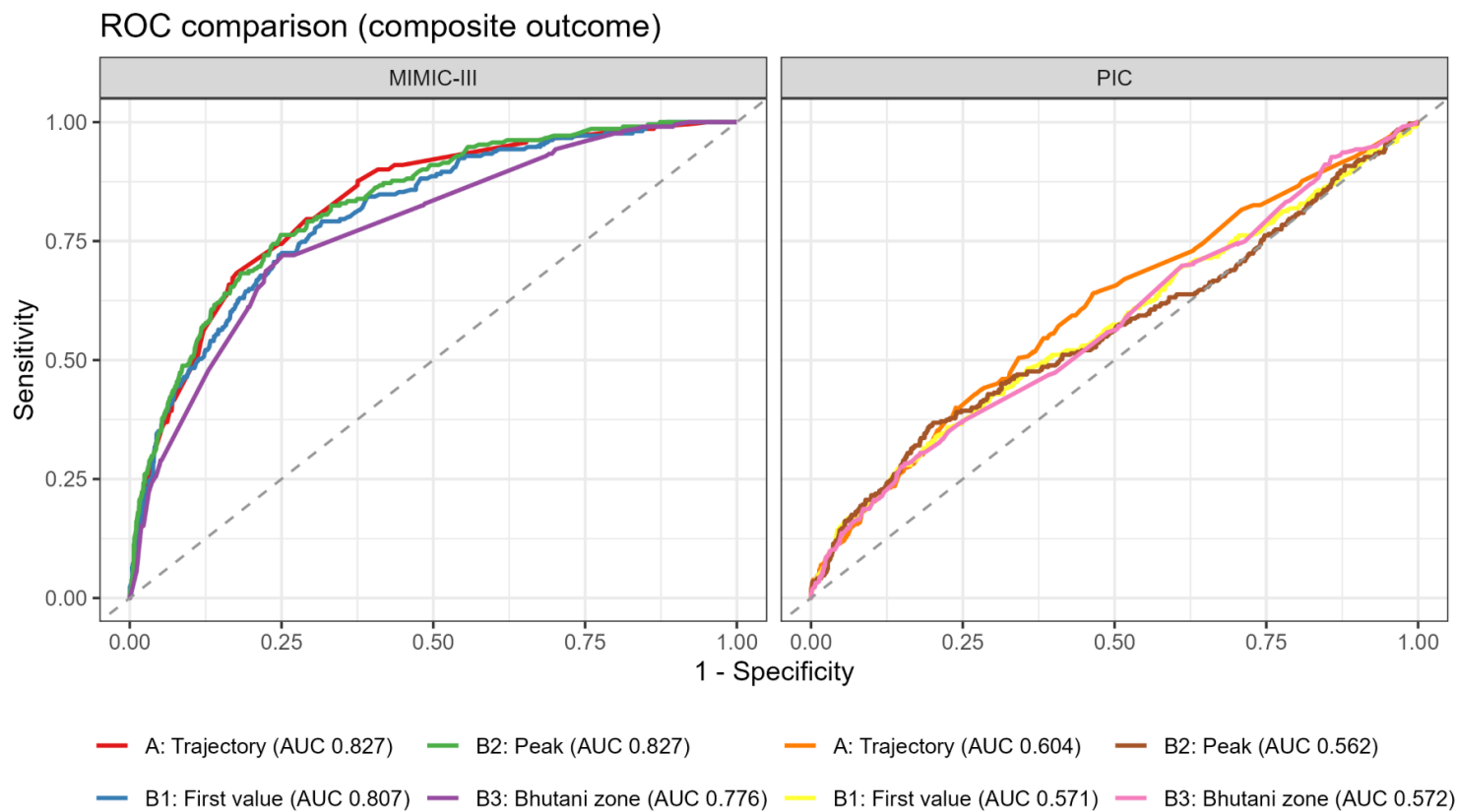


Figure 4

Receiver operating characteristic (ROC) curves comparing the trajectory model (A) with single-point models (B1, first value; B2, peak; B3, Bhutani risk zone) for the composite outcome in (a) MIMIC-III and (b) PIC.

Decision curve analysis (composite outcome): trajectory model A vs single-point models B1-B3

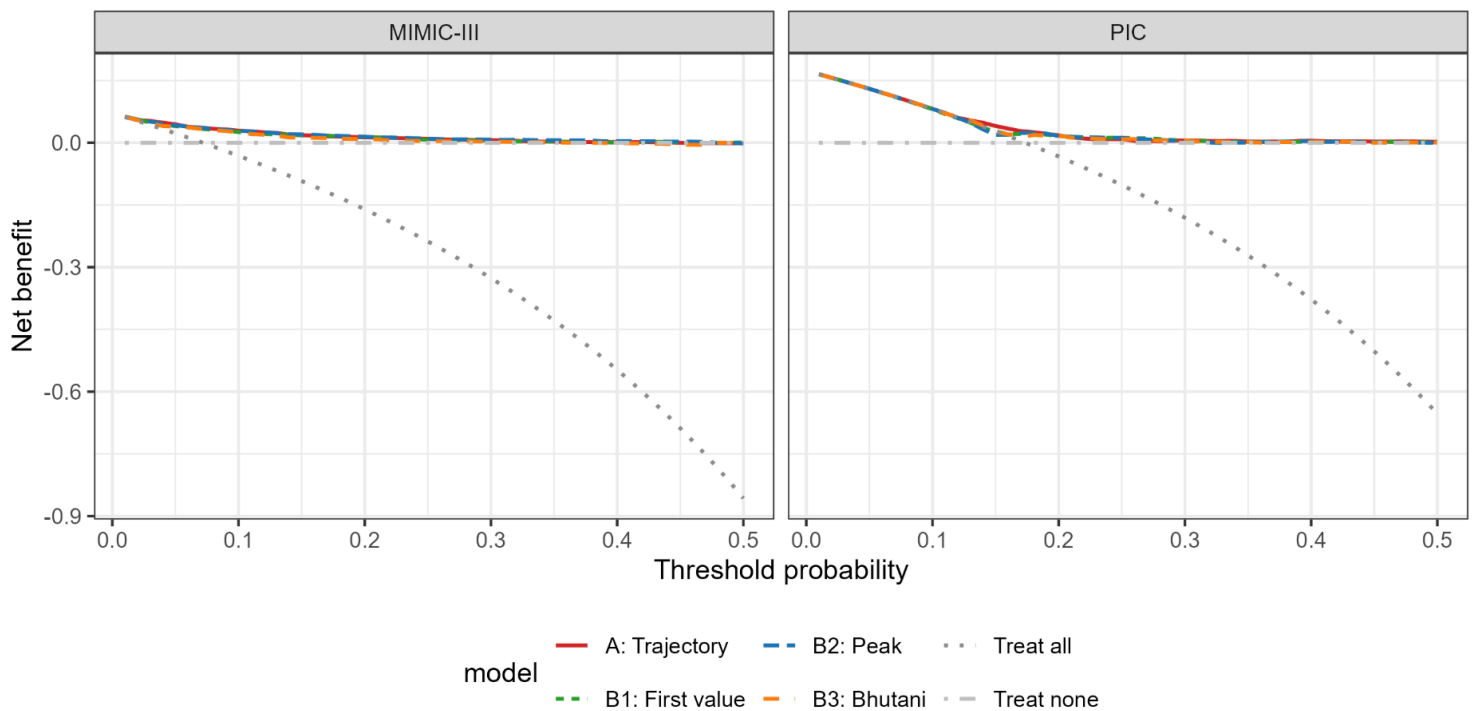


Figure 5

Decision curve analysis for the composite outcome in MIMIC-III and PIC: net benefit of the trajectory model (A) versus single-point models (B1, first value; B2, peak; B3, Bhutani risk zone) across threshold probabilities, with “treat all” and “treat none” reference strategies.

Supplementary Files

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

- [SupplementaryInformation.docx](#)