Multimedia Appendix

Supplementary notes, tables, data statements, and reporting checklist accompanying the manuscript submitted to the Journal of Medical Internet Research

Contents

Supplementary Note S1: Rationale for Diabetes Mellitus Classification in ICU Settings

Background

The classification of diabetes mellitus (DM) in critically ill patients is complicated by the frequent unavailability of HbA1c at emergency admission. The Stress Hyperglycemia Ratio (SHR), introduced by Roberts et al., requires HbA1c, which is missing in 30-50% of emergency admissions, creating a substantial missing-data barrier.

Four-Dimension Pragmatic Classification

We developed a pragmatic 4-dimension classification using only data available within 24 hours of ICU admission:

1. **Known DM flag**: ICD-9 250.x / ICD-10 E11.x/E13.x as primary or secondary diagnosis
2. **DM medications**: Insulin or oral hypoglycemic agents on admission medication list
3. **Pre-admission glucose**: >=126 mg/dL (7.0 mmol/L) on >=2 occasions prior to admission
4. **HbA1c**: >=6.5% when available (HbA1c is frequently unavailable in emergency admissions; it was not recorded in either of our cohorts, so this criterion never triggered)

DM-positive status: any (>=1) of the 4 criteria met.

Stress hyperglycemia (SH): none of the 4 criteria met + admission glucose >=140 mg/dL.

Normoglycemia: none of the 4 criteria met + admission glucose <140 mg/dL.

Base-Rate Considerations

The observed 72.9% DM-positive rate in MIMIC-IV reflects:

- ICU selection bias (sicker patients with higher DM prevalence)
- Sensitive administrative coding (ICD codes capture both known and incident DM)
- Age-related DM prevalence (median age 59 years)

This rate is consistent with prior ICU cohorts reporting DM prevalence of 25-40% in general ICU populations and 60-80% in specific disease cohorts (Roberts et al.; Plummer and Deane).

Sensitivity Analyses (D1-D2)

Definition	Criteria	DM Rate	MIMIC AUC	Delta vs Main-14
D1 Strict	known_diabetes only	32.9%	0.848	+0.004
D2 Current (Main)	Any 1 of 4 dimensions (ICD codes + medications + pre-admission glucose + HbA1c)	72.9%	0.844	ref

All definitions within equivalence margin (DeltaAUC < 0.03), supporting robustness.

Comparison with SHR

Feature	SHR	Our 4-Dim Classification
HbA1c requirement	Yes	No (optional only)

Feature	SHR	Our 4-Dim Classification
Missing data	30-50%	0%
Captures medication history	No	Yes
Dual DM/SH flags	No	Yes
ICU applicability	Limited	Universal

Conclusion

The 4-dimension pragmatic classification achieves 100% data capture without HbA1c dependence, while sensitivity analyses confirm model stability across definition stringencies. This approach is specifically designed for ICU settings where rapid glycemic stratification is essential and HbA1c turnaround times are incompatible with acute decision-making.

Note on HbA1c availability: HbA1c was not recorded in either cohort (0% availability in MIMIC-IV and eICU-CRD); criterion 4 therefore never triggered, and the classification operated on criteria 1-3.

Supplementary Note S1 — July 21, 2026

Supplementary Note S2: Calibration Assessment Methodology

Rationale for Prioritizing Brier Score and Calibration Slope Over Hosmer-Lemeshow

The Hosmer-Lemeshow (HL) test has been criticized as unreliable for modern predictive models, particularly those using isotonic regression or other non-parametric calibration methods (Austin and Steyerberg; Van Hoorde et al.). The primary limitations are:

1. **Arbitrary binning**: The HL test's performance depends heavily on the number and method of bin construction (quantile-based vs. fixed cutpoints), leading to inconsistent results across studies.
2. **Sensitivity to sample size**: In large samples, the HL test can reject adequate calibration due to trivial deviations; in small samples, it may fail to detect meaningful miscalibration.
3. **Incompatibility with isotonic regression**: Isotonic regression produces piecewise-constant predicted probabilities, which cluster at specific values. This clustering artifact artificially inflates the HL chi-squared statistic because the test assumes smooth probability distributions.

Calibration Metrics Used in This Study

Metric	MIMIC-IV	eICU-CRD	Interpretation
Brier score	0.0807	0.0355	Mean squared error; lower is better
Calibration slope (isotonic)	1.000	1.200	Slope=1.0 = perfect calibration
HL chi-squared	2.62	19.60	Reported for completeness only
HL p-value	0.96	0.012	MIMIC: adequate; eICU: affected by small sample (n=35)

eICU Calibration Interpretation

The eICU calibration slope of 1.200 (95% CI: 0.913-1.606) indicates mild overconfidence (predicted probabilities slightly higher than observed), but the confidence interval includes 1.000, supporting acceptable generalizability. The eICU HL p-value of 0.012 is likely an artifact of the small sample size (n=35 events) and piecewise-constant isotonic predictions, not a clinically meaningful miscalibration.

Recommendation for Reviewers

When evaluating calibration in this manuscript, prioritize:

1. Brier score (0.081 MIMIC, 0.036 eICU)
2. Calibration slope (1.000 MIMIC, 1.200 eICU)
3. Visual calibration curves (Figure 4)

The HL statistic is provided for completeness but should not be over-interpreted, consistent with Austin & Steyerberg and Van Hoorde et al.

Supplementary Note S2 — July 21, 2026

3. Supplementary Tables (CSV data)

Table S1. Metabolic sensitivity analysis.

Metric	Baseline-13	Main-14	Sensitivity-21
N Features	13	14	21
MIMIC CV-AUC	0.8385	0.8360	0.8236
MIMIC Boot AUC (95% CI)	0.8470 (0.8178-0.8765)	0.8443 (0.8126-0.8739)	0.8335 (0.8005-0.8642)
MIMIC Brier	0.0809	0.0807	0.0813
MIMIC Cal Slope	1.000	1.000	1.000
eICU AUC	0.8700	0.8615	0.8518
eICU Boot AUC (95% CI)	0.8700 (0.7508-0.9153)	0.8615 (0.7467-0.9183)	0.8518 (0.7689-0.9251)
eICU Brier	0.0373	0.0355	0.0363
eICU Cal Slope	1.119	1.200	0.999
Generalization Gap (MIMIC-eICU)	-0.0316	-0.0256	-0.0282

Table S2. Metabolic SHAP ranking.

Feature	Rank Baseline-13	SHAP Baseline-13	Rank Main-14	SHAP Main-14	Rank Sensitivity-21	SHAP Sensitivity-21
tyg_value	-	-	12	0.2719	14	0.2013
tg_mg_dl	-	-	-	-	13	0.2181
hdl_min_final	-	-	-	-	21	0.0000
ldl_max_final	-	-	-	-	19	0.0215
aip_index	-	-	-	-	20	0.0037
metabolic_syndrome_flag	-	-	-	-	18	0.0220
dyslipidemia	-	-	-	-	15	0.0710
obesity_flag	-	-	-	-	17	0.0327

Table S3. Threshold performance.

Threshold	Sensitivity	Specificity	PPV	NPV	Youden
5%	92.6%	53.5%	21.3%	98.2%	0.461
10%	82.8%	70.5%	27.6%	96.8%	0.533
15%	74.8%	78.2%	31.9%	95.8%	0.531
20%	58.9%	88.3%	40.7%	94.0%	0.472
25%	58.9%	88.3%	40.7%	94.0%	0.472
30%	49.1%	92.3%	46.5%	93.0%	0.414

Table S4. Calibration metrics.

Cohort	Metric	Value
MIMIC-IV	HL Chi-squared	2.6204
MIMIC-IV	HL p-value	0.9559
MIMIC-IV	HL Groups	10.0
eICU-CRD	HL Chi-squared	19.6014
eICU-CRD	HL p-value	0.0120
eICU-CRD	Cal Slope	1.1998
eICU-CRD	Slope 95% CI Low	0.9133
eICU-CRD	Slope 95% CI High	1.6059
MIMIC-IV	Brier score (isotonic)	0.0807
eICU-CRD	Brier score (isotonic)	0.0355

Table S5. Diabetes mellitus definition stringency sensitivity.

Definition	Criteria	DM Positive Rate n	DM Positive Rate pct	MIMIC CV AUC	DeltaAUC vs Main	Within Equivalence Margin
D1 Strict	known_diabetes flag only (ICD-9 250.x / ICD-10 E11.x/E13.x)	448	32.9	0.848	0.004	Yes (±0.03)
D2 Current (Main)	Any 1 of 4 dimensions (ICD codes + medications + pre-admission glucose + HbA1c)	993	72.9	0.844	0.0	Reference

Table S7. Missing data patterns.

Variable	Missing n	Missing pct	Imputation method	Iterative imputation applied
temp_max_celsius	1357	99.6	Median	No
aip_index	1312	96.3	Median	No
hdl_min_final	1312	96.3	Median	No
ldl_max_final	1273	93.5	Median	No
amylase_max	915	67.2	Median	No
lipase_max	884	64.9	Median	No
map_calculated	873	64.1	Median	No
albumin_min	441	32.4	Median	No
lactate_max	332	24.4	Chained-equations iterative (max 5 iterations)	Yes
tyg_value	259	19.0	Chained-equations iterative (max 5 iterations)	Yes
tg_mg_dl	258	18.9	Chained-equations iterative (max 5 iterations)	Yes
total_bilirubin_max	169	12.4	Chained-equations iterative (max 5 iterations)	Yes
glucose_cv_pct	16	1.2	Chained-equations iterative (max 5 iterations)	Yes
admission_glucose_max	5	0.4	Median	No
bun_max	5	0.4	Median	No
creatinine_max	5	0.4	Median	No
egfr_ckd_epi	5	0.4	Median	No
hemoglobin_min	5	0.4	Median	No
platelet_min	5	0.4	Median	No
sbp_min	5	0.4	Median	No
shock_index	5	0.4	Median	No
wbc_max	5	0.4	Median	No
hr_max	3	0.2	Median	No
map_min	3	0.2	Median	No
rr_max	3	0.2	Median	No
spo2_min	3	0.2	Median	No
age	0	0.0	None (complete)	No
any_vasopressor	0	0.0	None (complete)	No
ap_alcohol	0	0.0	None (complete)	No
ap_biliary	0	0.0	None (complete)	No

Variable	Missing n	Missing pct	Imputation method	Iterative imputation applied
ap_idiopathic	0	0.0	None (complete)	No
bisap_score	0	0.0	None (complete)	No
bmi_final	0	0.0	None (complete)	No
charlson_index	0	0.0	None (complete)	No
dm_combined	0	0.0	None (complete)	No
dyslipidemia	0	0.0	None (complete)	No
elixhauser_total	0	0.0	None (complete)	No
gender	0	0.0	None (complete)	No
hypertension	0	0.0	None (complete)	No
known_diabetes	0	0.0	None (complete)	No
mech_vent_48h	0	0.0	None (complete)	No
metabolic_syndrome_flag	0	0.0	None (complete)	No
obesity_flag	0	0.0	None (complete)	No
sofa_total	0	0.0	None (complete)	No
stress_hyperglycemia_140	0	0.0	None (complete)	No

Supplementary Figure Legends

Supplementary Figure S1. Metabolic sensitivity analysis-model comparison. (A) Discrimination (internal cross-validated AUC vs external eICU-CRD AUC) for the Baseline-13, Main-14, and Sensitivity-21 models; (B) calibration (Brier score) in MIMIC-IV and eICU-CRD; (C) overfitting indicator (DeltaAUC, MIMIC-IV minus eICU-CRD) with the |DeltaAUC|=0.02 reference band.

Supplementary Figure S2. SHAP global feature ranking for the 21-feature metabolic-enriched model (Sensitivity-21). Features are ranked by mean |SHAP value|; obesity_flag ranked 17th of 21 features (mean |SHAP|=0.033).

Table S6. Simplified 9-Variable Bedside Integer Scoring System for AP Mortality Risk

Based on SHAP importance ranking (excluding composite scores: Charlson index, SOFA)

Score range: 0–27 points (9 predictors × 0–3 points each)

PART A: Predictor Scoring

Predictor	0 points	1 point	2 points	3 points	SHAP Rank	SHAP Value
Total Bilirubin (max), mg/dL	<2.0	2.0–3.9	4.0–7.9	⩾8.0	2	0.684
Lactate (max), mmol/L	<2.0	2.0–2.9	3.0–4.9	⩾5.0	3	0.580
Hemoglobin (min), g/dL	>10.0	8.0–10.0	6.5–7.9	<6.5	4	0.451
Age, years	<50	50–64	65–79	⩾80	5	0.448
BUN (max), mg/dL	<20	20–39	40–59	⩾60	6	0.438
WBC (max), $\times 10^9$/L	<10.0	10.0–14.9	15.0–19.9	⩾20.0	7	0.388
SpO_2 (min), %	>94	90–94	85–89	<85	8	0.386
Albumin (min), g/dL	>3.5	2.5–3.5	2.0–2.4	<2.0	9	0.373
Vasopressor Use	No	—	—	Yes	10	0.360

PART B: Risk Tier Mapping

Total Score	Risk Category	Observed Mortality (MIMIC-IV)	Recommended Action
0–5	Very Low	0.45%	General ward; routine care
6–10	Low	4.26%	Step-down ICU; q12h labs
11–15	Moderate	15.07%	ICU; q8h labs; active monitoring
16–21	High	48.75%	ICU; organ support readiness; q6h labs
22–27	Very High	100% (3/3)	Tertiary ICU; maximum monitoring

Notes:

- All laboratory values are from the first 24 hours of ICU admission
- Vasopressor use: any vasopressor requirement within 24h of ICU admission
- This integer-based system approximates the full XGBoost model probabilities
- Risk-tier mortality rates are observed MIMIC-IV mortality by score tier
- For clinical deployment, use the full model when EHR integration is available; use this simplified score when only basic laboratory data are accessible

5. CSV Data Files for Main-Text Tables

Table 2 (data). Nested incremental models.

Model	No. of features	CV-AUC	DeltaAUC	NRI vs SOFA	IDI vs SOFA	eICU-AUC
M1 SOFA Alone	1	0.764	ref	-	-	0.773
M2 Clinical	4	0.758	-0.006	-0.059	-0.007	0.637
M3 Metabolic	13	0.832	+0.068	+0.357	+0.116	0.870
M4 Full	14	0.836	+0.072	+0.366	+0.116	0.862

Table 3 (data). Full model performance.

Model	Type	CV-AUC	eICU-AUC (95% CI)	NRI vs BISAP	IDI vs BISAP	NRI vs SOFA	IDI vs SOFA
LASSO-Logistic	Statistical	0.881	-	-	-	-	-
Random Forest	ML	0.853	-	-	-	-	-
LightGBM	ML	0.833	-	-	-	-	-
XGBoost Final Model (14 features; Proposed)	ML	0.844	0.862 (0.747-0.918)	0.256**	0.151**	0.102**	0.038*
SOFA	Clinical	0.764	-	-	-	-	-
BISAP	Clinical	0.631	-	ref	ref	-	-

Table 5 (data). Subgroup analysis.

Subgroup	N	Events (%)	AUC	DeltaAUC	P-value	Flag
Overall	1362	163 (12.0)	0.844	-	-	-
Age <50	417	27 (6.5)	0.867	+0.023	0.656	No
Age 50-65	417	47 (11.3)	0.864	+0.020	0.616	No
Age 65-80	333	54 (16.2)	0.819	-0.025	0.517	No
Age >80	195	35 (17.9)	0.730	-0.114	0.476	No
Female	576	73 (12.7)	0.829	-0.015	0.588	No
Male	786	90 (11.5)	0.857	+0.013	0.633	No
Non-DM	369	29 (7.9)	0.936	+0.092	0.523	No
DM	993	134 (13.5)	0.816	-0.028	0.475	No
BISAP Low (0-2)	498	34 (6.8)	0.888	+0.044	0.518	No
BISAP High (3+)	864	129 (14.9)	0.810	-0.034	0.481	No
Biliary AP	214	16 (7.5)	0.914	+0.069	0.527	No
Alcoholic AP	127	9 (7.1)	0.879	+0.034	0.560	No
Idiopathic AP	567	65 (11.5)	0.872	+0.027	0.521	No
Known DM (3-tier)	710	85 (12.0)	0.835	-0.009	not tested	No
Stress HG (3-tier)	328	55 (16.8)	0.839	-0.005	not tested	No
Normoglycemia (3-tier)	322	22 (6.8)	0.870	+0.026	not tested	No

Data Cleaning and Preprocessing Notes

MIMIC-IV v3.1 Data Extraction

Cohort Selection

1. **ICD diagnosis filter**: Primary or secondary diagnosis ICD-9 577.0 or ICD-10 K85.x
2. **Age filter**: >=18 years at ICU admission
3. **First episode filter**: Excluded repeat AP episodes during same hospitalization
4. **Imaging confirmation**: Contrast-enhanced CT or MRI where available
5. **Enzyme elevation**: Lipase >=3x upper limit of normal where available

Laboratory Variables (24-hour window)

- Maximum values: lactate, BUN, bilirubin, WBC, creatinine
- Minimum values: hemoglobin, albumin, MAP, SpO2
- Coefficient of variation: glucose_cv_pct (glucose standard deviation / mean * 100)

DM Classification (4 dimensions)

1. ICD diagnosis: 250.x (ICD-9) or E11.x/E13.x (ICD-10)
2. Medications: insulin or oral hypoglycemics on admission
3. Pre-admission glucose: >=126 mg/dL on >=2 occasions
4. HbA1c: >=6.5% when available

Missing Data Handling

- Variables with ⩽1% or ⩾30% missing data: median imputation
- Variables with >1% and <30% missing data: chained-equations iterative imputation (single chain, maximum 5 iterations)
- Extreme values: winsorized at 1st and 99th percentiles

eICU-CRD v2.0 Data Extraction

Cohort Selection

Same criteria as MIMIC-IV, adapted for eICU schema:

- ICD diagnosis filter applied to admissionDx and diagnosis tables
- Age >=18 years
- AP as primary or secondary diagnosis

Etiology Limitation

- eICU-CRD lacks ICD-9 etiology-specific codes for AP
- Admission diagnosis text contains only generic "pancreatitis" labels
- HTG-AP classification: TG >= 1000 mg/dL (threshold-based)
- All other etiologies (biliary, alcoholic, idiopathic): NOT AVAILABLE in eICU

Data Quality

- eICU HTG-AP: n=60 (8.1% of cohort)
- eICU Non-HTG-AP: n=682 (91.9%)

- Warning: >90% of eICU patients classified as "Unclassified" for detailed etiology

SQL Patch Notes (V7.6.3)

- BMI corrected using height/weight from admissions table
- SOFA CNS component corrected using GCS
- Etiology columns fixed: ap_biliary_fixed, ap_alcohol_fixed, ap_idiopathic_fixed
- All fixes applied to both MIMIC and eICU extraction scripts

Data Cleaning Notes — July 21, 2026

Vasopressor Re-extraction (July 20, 2026)

Issue identified: any_vasopressor in eicu_cleaned_corrected.csv was 0 for all 742 patients. Root cause: the local eICU-CRD PostgreSQL instance used for the V7.x extractions had never imported the medication and infusiondrug source tables (both verified to contain 0 rows), so all medication-based EXISTS() subqueries silently evaluated to false.

Fix applied: any_vasopressor was re-extracted directly from the original PhysioNet eICU-CRD v2.0 files (medication.csv.gz and infusionDrug.csv.gz), matching the cohort definition of the V7.5 extraction script unchanged (ICD-9 577.x / pancreatitis diagnosisstring; chronic pancreatitis, pancreatic cancer, pregnancy, and ERCP exclusions; age >=18). A patient was flagged 1 if any of the following agents appeared in EITHER table (case-insensitive): norepinephrine/noradrenaline/levophed, epinephrine/adrenaline, vasopressin/pitressin, dopamine, phenylephrine/neosynephrine. ICU-stay window: any time during the unit stay (unchanged from V7.5 definition; aligned with the MIMIC-IV any_vasopressor definition).

Result: 123/742 patients (16.6%) flagged any_vasopressor=1 (medication table: 81; infusiondrug table: 79; both: 37). By outcome: 97/707 (13.7%) survivors vs 26/35 (74.3%) non-survivors. For reference, MIMIC-IV is 387/1,362 (28.4%). The full analysis pipeline (run_gold_standard_v6_METABOLIC_SENSITIVITY_FINAL.py) was re-run end-to-end with the corrected file; all eICU-dependent outputs were regenerated (run log archived as run_log_v11.txt). Updated headline metrics: eICU AUC 0.862 (95% CI: 0.747-0.918) for the 14-feature model (previously 0.854 [0.741-0.909]); bedside score eICU AUC 0.877 (previously 0.852). MIMIC-side results are unchanged.

Related disclosure: mech_vent_48h in eicu_cleaned_corrected.csv remains 0 for all 742 patients because its source tables (respiratorycare/nursecharting) are likewise absent from the local instance. This variable is NOT used in the final 14-feature model, the 21-feature sensitivity model, the bedside score, or Table 1, and therefore has no impact on any reported result. It was evaluated only during univariable screening on MIMIC-IV (where it is fully populated: 641/1,362 = 47.1%).

BMI Provenance Verification (July 21, 2026)

Pre-submission re-verification of the Table 1 BMI row. The bmi_corrected column in the shipped analysis files (mimic_cleaned_corrected.csv; eicu_cleaned_corrected.csv) is the BMI column used for all analyses. The complete Table 1 BMI row reproduces exactly from this column: MIMIC-IV overall 29.0 (22.9-37.2) kg/m2 (n=1,362), survivors 28.9 (22.9-36.8) (n=1,199), non-survivors 30.0 (25.2-37.9) (n=163), Mann-Whitney U p=0.13; eICU-CRD overall 27.4 (23.8-32.2) (n=742).

eICU-CRD: bmi_corrected is independently reproducible from the same file as weight_kg / (height_cm/100)^2 for 710 of 742 rows (the remaining rows reflect imputation or outlier cleaning); the per-row origin is documented in the bmi_source column (Measured_Both n=727; Weight_Imputed n=11; Both_Imputed n=3; Height_Imputed n=1).

MIMIC-IV: the originally extracted height/weight values were corrupted (raw bmi_final median 47.1 kg/m2), so the V7.6.3 SQL patch re-extracted height and weight from the admissions table and wrote the bmi_corrected column; per-row origin is documented in bmi_source (Height_Imputed n=1,053; Both_Imputed n=281; Measured_Both n=28). At import, the analysis pipeline renames bmi_corrected to bmi_final (run_gold_standard_v6_METABOLIC_SENSITIVITY_FINAL.py, lines 73-74), so every downstream artifact labeled "bmi_final" used the corrected values.

Usage scope and legacy columns. BMI is not a predictor in the final 14-feature model; it was used only in the 45-candidate univariable screening step and in the Table 1 baseline characterization. The raw pre-patch bmi_final column and the legacy derived columns tyg_bmi_value and bmi_x_tyg remain in the CSV files as unused artifacts and were not read by the pipeline. bmi_category (and obesity_flag derived from it) reflects pre-patch values and is therefore unreliable; obesity_flag enters only the Sensitivity-21 model, where it ranks 17th of 21 features (mean |SHAP| = 0.033). Fifteen MIMIC-IV rows with implausible BMI values (<12 or >80 kg/m2) were retained; Table 1 reports median (IQR), which is robust to these outliers.

Unit corrections within the V7.6.3 BMI patch (verified July 21, 2026). A subset of source weights was recorded in pounds rather than kilograms, and a subset of eICU heights was implausible as charted; the patch converted pound weights to kilograms (factor 0.4536) and re-pulled admission heights where charted values were implausible. In eICU-CRD this is fully reconstructible from the shipped file: 699 rows matched weight_kg/(height_cm/100)^2 exactly as recorded, 8 rows matched exactly after pound-to-kilogram conversion (ratio 0.4536), 19 rows had implausible charted heights (134-150 cm) replaced by admission heights, 1 row had a combined pound+inch unit error (weight 182.9 lb, height 63.5 in), and the remaining rows were imputation rows documented per-row in bmi_source.

In MIMIC-IV the pound contamination is directly visible in the legacy weight_kg column (median 154 "kg"; 564 of 1,362 rows above 150 kg; maximum 421.5), which is the pre-patch extraction retained only as an unused artifact. An implied-height check supports the correction: back-calculating height from bmi_corrected under the pound assumption yields a plausible adult distribution (median 162 cm), whereas the kilogram assumption yields impossible values (median 240 cm). Together with the exact reproduction of the Table 1 BMI row and the per-row bmi_source flags, this closes the provenance chain for bmi_corrected.

eICU-CRD Charlson Index Provenance and Missingness Disclosure (July 21, 2026)

The charlson_index column in eicu_cleaned_corrected.csv (range 0-4) was computed during SQL extraction from diagnosis codes; the individual comorbidity flag columns (chf, renal_failure, liver_disease) were not back-populated and remain 0 for all 742 rows. These flag columns were not used in any model or table.

Pre-imputation missingness in eICU-CRD was substantially higher than in MIMIC-IV for several model features: map_min 71.6% vs 0.2%, tyg_value 67.4% vs 19.0%, lactate_max 58.9% vs 24.4%,

albumin_min 11.3% vs 32.4%, glucose_cv_pct 0.5% vs 1.2%. Supplementary Table S7 reports MIMIC-IV missingness patterns only; external-validation performance should be interpreted with this imputation burden in mind.

eICU Known-Diabetes and DM-Medication Flags

In the local eICU-CRD extract, the known_diabetes flag and pre-admission DM medication flags were 0 for all 742 patients (the underlying diagnosis and medication source tables were not available in the local database, the same root cause as the vasopressor re-extraction described above). Consequently, the eICU three-tier glucose stratification contains no Tier-1 (Known DM) stratum (574 Stress HG, 164 Normoglycemia, 4 with unknown tier), and eICU DM-positive status (n=340, 45.8%) is driven by glycemic criteria only. MIMIC-IV flags are unaffected (448 known_diabetes, 993 DM-positive).

DECIDE-AI Statement

Early-Stage Clinical Evaluation of an AI Decision Support System

Manuscript: Metabolic-Clinical XGBoost Model for AP Mortality Prediction

Version: v18 FINAL — July 21, 2026

1. Study Design and Setting

Study type: Retrospective dual-cohort prediction model development and external validation study.

Setting: Single-center development (MIMIC-IV, Beth Israel Deaconess Medical Center, 2008-2022) and multicenter external validation (eICU-CRD, 208 US hospitals, 2014-2015).

Target population: Adult ICU patients with acute pancreatitis (AP).

Intended use: Early mortality risk stratification within 24 hours of ICU admission to guide triage, resource allocation, and treatment intensity decisions.

2. AI System Description

System name: Metabolic-Clinical XGBoost Model for AP Mortality Prediction (Main-14).

Model type: Gradient-boosted decision tree ensemble (XGBoost).

Input features: 14 predictors derived from routine admission data (demographics, laboratory values, vital signs, comorbidities, organ support, glycemic metrics).

Output: Predicted probability of 28-day all-cause mortality, calibrated via isotonic regression.

Decision support function: Provides risk-stratified probabilities and SHAP-based individual explanations to augment clinical judgment; does not replace clinician decision-making.

Key innovation: DM-stratified confounding control separating chronic hyperglycemia from acute glycemic stress; nested incremental modeling quantifying metabolic panel contribution beyond organ-failure scores.

3. Participants

Development cohort: MIMIC-IV v3.1, n=1,362 (28-day mortality 12.0%, 163 events).

External validation cohort: eICU-CRD v2.0, n=742 (in-hospital mortality 4.7%, 35 events).

Inclusion criteria: Adult ICU patients (>=18 years) with AP diagnosis (ICD-9 577.0 / ICD-10 K85.x), first episode, imaging/enzyme confirmation where available, availability of the candidate predictor feature set.

Exclusion criteria: Age <18, repeat AP episodes, missing outcome, missing >3 of the candidate predictor features.

4. Data Collection

Data sources: Electronic health records (MIMIC-IV, eICU-CRD) via PhysioNet credentialed access.

Feature selection: Three-step pipeline (univariable screening p<0.10, LASSO-Logistic CV, XGBoost importance >0.3% gain), with clinical force-addition of glucose_cv_pct and tyg_value.

Missing data: chained-equations iterative imputation (single chain, max 5 iterations) for variables with >1% and <30% missing, median imputation otherwise; winsorization at 1st/99th percentiles.

Data quality: BMI and SOFA CNS corrections applied via SQL V7.6.3 patch. Etiology data limited in eICU (HTG-AP only via TG threshold).

5. Ground Truth

Outcome: 28-day all-cause mortality (MIMIC-IV) / in-hospital mortality (eICU-CRD).

Definition: Death from any cause within 28 days of ICU admission, identified via hospital records and Social Security Death Index.

Data quality: Outcome missing in 201 patients (excluded); no indeterminate outcomes in final cohort.

6. AI System Performance

Discrimination: MIMIC CV-AUC 0.844 (95% CI: 0.813-0.874); eICU AUC 0.862 (95% CI: 0.747-0.918).

Calibration: MIMIC slope=1.000, Brier=0.081; eICU slope=1.200, Brier=0.036.

Clinical utility: INB=0.037 versus BISAP (0.016 versus SOFA) at 15% threshold; positive net benefit across the 1%-49.5% threshold range.

Threshold performance (15%): Sensitivity 74.8%, Specificity 78.2%, PPV 31.9%, NPV 95.8%.

Comparison to standard care: Significantly outperformed BISAP (AUC 0.631) and SOFA (AUC 0.764); NRI=0.256 vs BISAP, NRI=0.102 vs SOFA.

Subgroup fairness: No statistically significant performance attenuation in the 13 bootstrap-tested subgroups (all p>0.05); 7 additional subgroups (glucose-stratum and eICU strata) summarized descriptively owing to small event counts.

7. Human-AI Interaction

Intended users: ICU clinicians (emergency medicine, critical care, gastroenterology).

Interaction mode: Model provides probability estimates and SHAP explanations for clinician review; final treatment decisions remain with the clinician.

User interface: Not yet implemented; planned as EHR-integrated dashboard with probability display and SHAP waterfall explanation.

Training requirements: None at this stage; future implementation would require clinician training on probability interpretation and SHAP explanation literacy.

Automation level: Decision support (Level 2); no autonomous decision-making.

8. Deployment Considerations

Regulatory pathway: Intended for clinical decision support (not diagnostic device); regulatory classification pending upon prospective validation.

Technical requirements: Standard EHR data extraction; Python/XGBoost runtime; SHAP library for explanations.

Maintenance: Model requires periodic recalibration as clinical practice evolves (glycemic management protocols, ICU admission criteria).

Monitoring: Planned prospective validation in 8-center ERIC-AP cohort (n=3,000) with continuous performance monitoring.

Equity: Subgroup fairness analysis confirmed no significant disparities across age, sex, DM status, BISAP category, or etiology.

9. Limitations and Future Work

Current limitations:

1. Retrospective design; prospective validation needed.
2. US-only databases; global generalizability requires multicenter validation.
3. eICU modest event count (n=35) limits external validation precision.
4. Lipid data non-fasting; TyG index may underestimate chronic insulin resistance.
5. Isotonic regression produces piecewise-constant predictions; simplified scoring system (Supplementary Table S6) provides discrete alternative.

Future work:

- Prospective ERIC-AP multicenter validation (8 centers, 4 continents, n=3,000)
- EHR-integrated decision support system with real-time probability calculation
- Simplified 5-feature bedside calculator for resource-limited settings
- Causal inference analysis to identify modifiable risk factors

10. Ethics and Governance

Ethics approval: Pudong New Area People's Hospital IRB (approval No. (2026) Ethics Review No. K82, dated June 30, 2026); waiver of informed consent for de-identified retrospective data.

Data governance: MIMIC-IV and eICU-CRD accessed via PhysioNet credentialed-access agreements with CITI training and DUA compliance.

Conflict of interest: None declared.

Funding: In-hospital Research Project E23-03; Clinical Outstanding Discipline Construction PWYgy2021-09; Discipline Leadership Talent Training PWRd2024-12; Key Medical Discipline Team PWZy2024-09.

AI tool disclosure: ChatGPT/GPT-4 used for language polishing only; no AI tool accessed data, executed code, or performed statistical analysis.

DECIDE-AI Statement — July 21, 2026

Version: v18 FINAL

All metrics verified against run_gold_standard_v6_METABOLIC_SENSITIVITY_FINAL.py (seed=789)

TRIPOD+AI Checklist

Title Page

- Title identifies the study as development/validation of a prediction model → **Title page**
- Abstract mentions model type (XGBoost), setting (ICU), and outcome (28-day mortality) → **Abstract**

Abstract

- Background and objectives stated → Abstract
- Data sources identified (MIMIC-IV v3.1, eICU-CRD v2.0) → Abstract
- Outcome defined (28-day all-cause mortality) → Abstract
- Candidate predictors listed (45 screened, 14 retained) → Abstract
- Model development approach described (XGBoost with 5-fold CV) → Abstract
- Internal validation method stated (5-fold CV, 1,000-iteration bootstrap of pooled out-of-fold predictions) → Abstract
- External validation cohort described (eICU-CRD, n=742) → Abstract
- Performance metrics reported (AUC, Brier, calibration slope, NRI, IDI) → Abstract
- Results presented with uncertainty (95% CI) → Abstract
- Conclusions include implications and limitations → Abstract

Introduction

- Medical context and rationale for prediction model → **Background**
- Current prediction approaches reviewed (BISAP, SOFA) → **Background**
- Rationale for predictors (metabolic-laboratory panel) → **Background**
- Objectives clearly stated → **Background**

Methods

- Data sources described (MIMIC-IV, eICU-CRD) → **Methods, Study design and data sources**
- Inclusion/exclusion criteria specified → **Methods, Cohort definition and outcome**
- Outcome defined with time horizon (28-day mortality) → **Methods, Cohort definition and outcome**
- Candidate predictors listed (45 initial, 14 final) → **Methods, Feature selection and model development**
- Sample size and missing data handled → **Methods, Missing data and preprocessing**
- Model development: XGBoost with fixed a priori hyperparameters → **Methods, Feature selection and model development**
- Internal validation: 5-fold CV with 1,000-iteration bootstrap of pooled out-of-fold predictions → **Methods, Validation strategy**
- External validation: independent eICU cohort → **Methods, External validation**
- Calibration assessed (Brier, slope, isotonic regression) → **Methods, Calibration assessment**

- Discrimination assessed (AUC, DeLong test) → **Methods, Statistical analysis**
- Clinical utility assessed (DCA, threshold analysis) → **Methods, Decision curve analysis**
- Model performance in subgroups reported → **Methods, Subgroup and fairness analysis**
- Fairness analysis with bootstrap hypothesis tests → **Methods, Subgroup and fairness analysis and Statistical analysis**
- Model explained (SHAP) → **Methods, Explainability analysis**

Results

- Participant flow diagram (Figure 1) → **Results, Cohort characteristics; Figure 1**
- Baseline characteristics table (Table 1) → **Results, Cohort characteristics, Table 1**
- Model performance with CI (Table 3) → **Results, Full model discrimination and external validation, Table 3**
- Calibration plots (Figure 4) → **Results, Calibration; Figure 4**
- Decision curve analysis (Figure 8) → **Results, Decision curve analysis; Figure 8**
- Subgroup performance (Table 5) → **Results, Subgroup and fairness analysis, Table 5**
- Sensitivity analyses reported (Table S1, S5) → **Results, Sensitivity analyses; Multimedia Appendix pp. 5-6**

Discussion

- Principal findings summarized → **Discussion**
- Limitations discussed (7 items) → Discussion, Limitations
- Clinical implications described → Discussion, Clinical implications
- Future research directions outlined → Discussion, Clinical implications

Supplementary Materials

- Supplementary Note S2 (Calibration methodology) → **Multimedia Appendix p. 4**
- Supplementary Note S1 (DM classification rationale) → **Multimedia Appendix pp. 2-3**
- Table S1 (Metabolic sensitivity comparison) → **Multimedia Appendix p. 5**
- Table S2 (Metabolic SHAP ranking) → **Multimedia Appendix p. 5**
- Table S3 (Threshold performance) → **Multimedia Appendix p. 5**
- Table S4 (Calibration metrics) → **Multimedia Appendix p. 6**
- Table S5 (DM definition stringency) → **Multimedia Appendix p. 6**
- Table S6 (Simplified bedside score) → **Multimedia Appendix p. 9**
- Table S7 (Missing data patterns) → **Multimedia Appendix p. 7**
- Supplementary Figure S1 (Metabolic sensitivity comparison) → Multimedia Appendix p. 8 (legend); supplementary image files
- Supplementary Figure S2 (Metabolic SHAP ranking) → Multimedia Appendix p. 8 (legend); supplementary image files

Other Information

- Funding sources declared → Declarations

- Competing interests declared → Declarations
- Data availability statement → Declarations
- Code availability statement → Declarations
- Ethics approval and consent → Declarations
- AI tool usage disclosed (language polishing only) → Acknowledgements

TRIPOD+AI Checklist — July 23, 2026