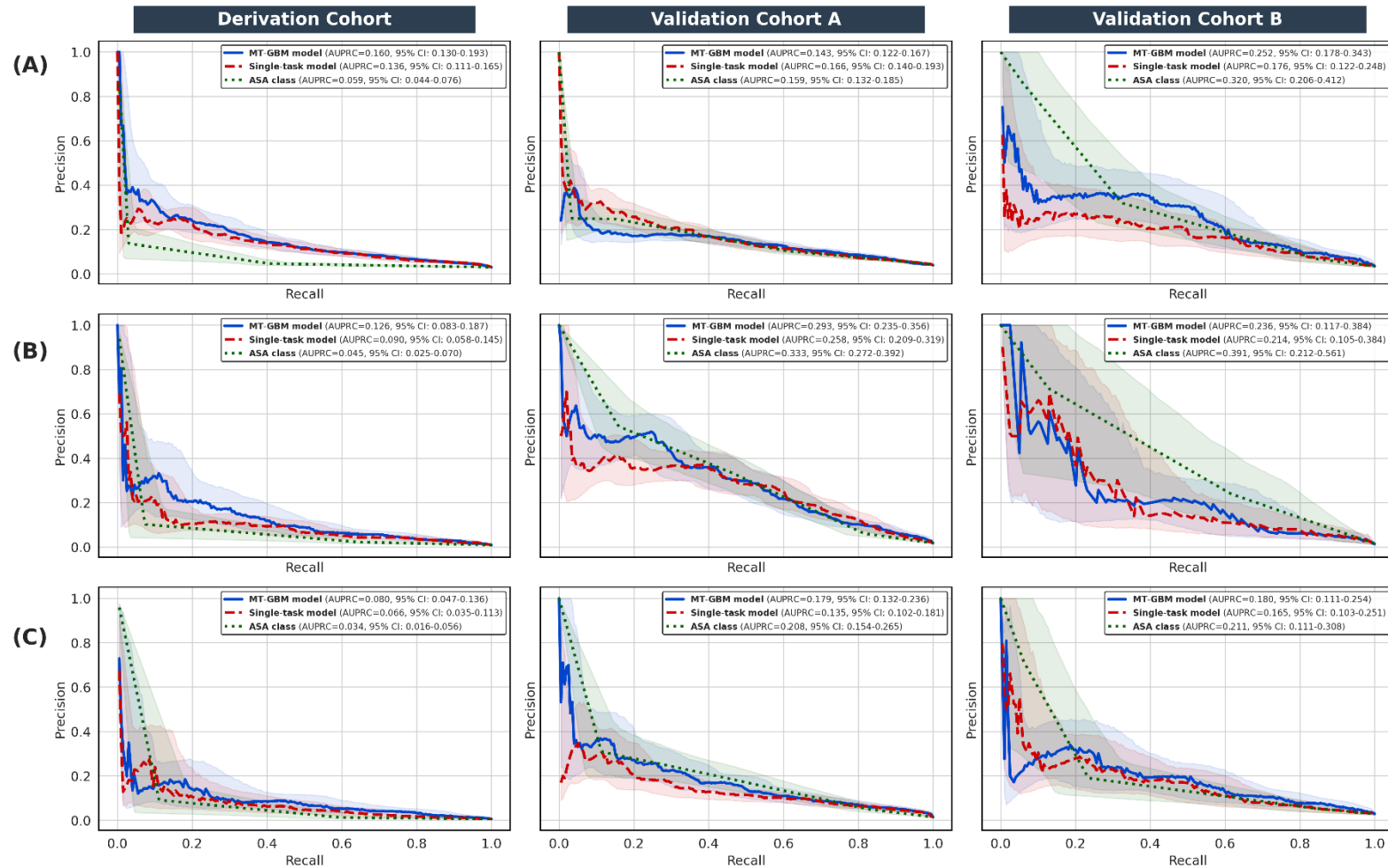
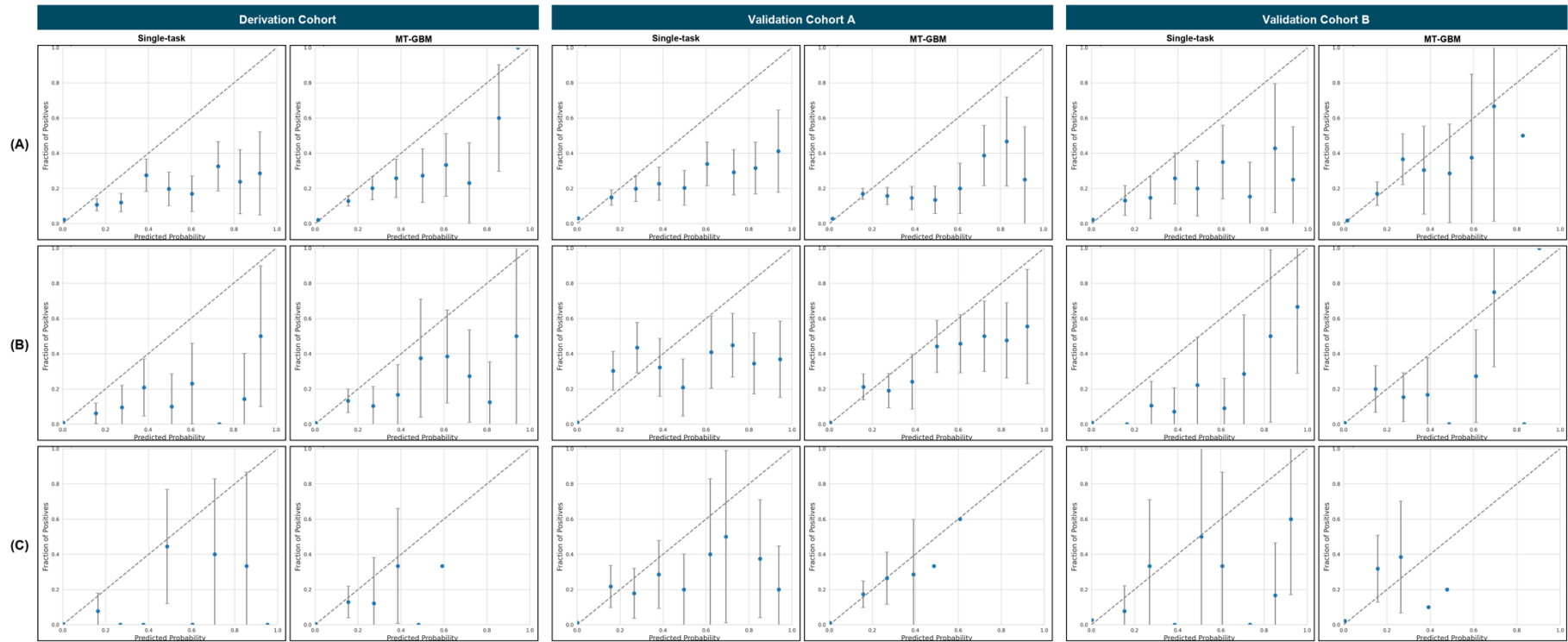


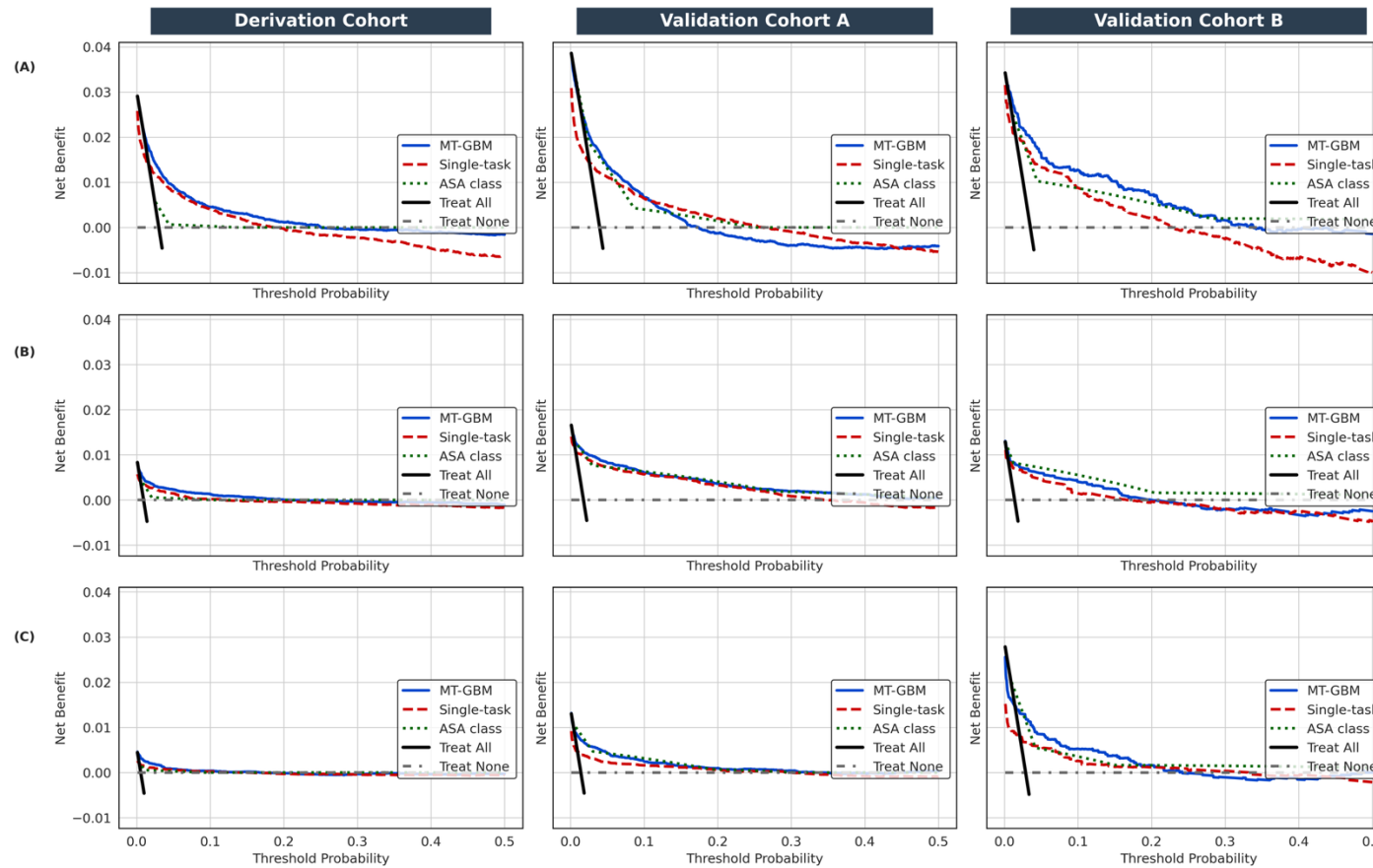
Supplementary Figure S1. Area under the precision-recall curves (AUPRC) comparing the performance of the MT-GBM and single-task models, and ASA class for predicting (A) acute kidney injury, (B) postoperative respiratory failure, and (C) in-hospital mortality across the derivation cohort and two validation cohorts. Shaded regions represent 95% confidence intervals calculated using bootstrap resampling (1,000 iterations). Differences in AUPRC between models were statistically compared using bootstrap paired tests. MT-GBM: multitask gradient boosting machine, ASA: American Society of Anesthesiologists physical status classification.



Supplementary Figure S2. Calibration plots for the derivation and validation cohorts. (A) Predictions for acute kidney injury, (B) Predictions for postoperative respiratory failure, and (C) Predictions for in-hospital mortality. The diagonal dashed line in each plot represents perfect calibration, where the predicted probability matches the observed frequency of positive outcomes. Each blue point shows the observed proportion of positives for a given range of predicted probabilities, and the vertical bars around each point indicate the 95% confidence interval for that estimate. MT-GBM: multitask gradient boosting machine.



Supplementary Figure S3. Decision Curve Analysis (DCA) across prediction tasks and cohorts. (A) Predictions for acute kidney injury, (B) Predictions for postoperative respiratory failure, and (C) Predictions for in-hospital mortality. Each panel compares the net benefit of four decision strategies across threshold probabilities: MT-GBM (multi-task gradient boosting model, blue solid line), Single-task model (red dashed line), ASA (American Society of Anesthesiologists physical status classification, green dotted line), Treat All (gray solid line), and Treat None (black dash-dot line). Columns represent the Derivation Cohort, Validation Cohort A, and Validation Cohort B from left to right. The net benefit is calculated based on true positives and false positives for each threshold, enabling comparison of clinical utility across models. MT-GBM: multitask gradient boosting machine, ASA: American Society of Anesthesiologists physical status classification.



Supplementary Table S1. Missing data in demographic and preoperative laboratory variables in the derivation and validation cohorts.

Characteristics	Derivation Cohort (n = 66,152)		Validation Cohort A (n = 13,285)		Validation Cohort B (n = 2813)	
	Missing data (n)	Missing data (%)	Missing data (n)	Missing data (%)	Missing data (n)	Missing data (%)
Age, years	0	0	0	0	0	0
Sex, n (%)	0	0	0	0	0	0
Body mass index, kg/m ²	1687	2.55	0	0	93	3.31
Emergency surgery, n (%)	0	0	0	0	0	0
Preoperative laboratory test						
Hemoglobin, g/dL	1913	2.89	0	0	0	0
White blood cell counts, 10 ³ /μL	4305	6.51	0	0	0	0
Platelet, 10 ³ /μL	1976	2.99	0	0	0	0
Blood urea nitrogen, mg/dL	2903	4.39	0	0	0	0
Serum creatinine, mg/dL	0	0	0	0	0	0
Serum glucose, mg/dL	4831	7.3	0	0	7	0.25
Albumin, mg/dL	1189	1.8	0	0	0	0
AST, IU/L	3177	4.8	0	0	0	0
ALT, IU/L	3183	4.81	0	0	0	0
Sodium, mmol/L	1531	2.31	0	0	1	0.04
Potassium, mmol/L	1515	2.29	0	0	1	0.04
Prothrombin time, INR	6889	10.41	0	0	1	0.04

AST: aspartate aminotransferase, ALT: alanine aminotransferase, INR: international normalized ratio.

Supplementary Table S2. Comparisons of AUPRC and F1-score between MT-GBM and single-task or ASA class prediction models for acute kidney injury, postoperative respiratory failure, and in-hospital mortality across derivation and validation cohorts.

Cohorts	Metrics	MT-GBM vs Single-task		MT-GBM vs ASA class	
		Mean Difference (95% CI)	<i>p</i> -value	Mean Difference (95% CI)	<i>p</i> -value
Acute kidney injury					
Derivation Cohort	AUPRC	0.024 (0.000 to 0.049)	0.027	0.101 (0.072 to 0.134)	<0.001
Derivation Cohort	F1-score	0.021 (-0.006 to 0.049)	0.060	0.164 (0.132 to 0.199)	<0.001
Validation Cohort A	AUPRC	-0.022 (-0.037 to -0.006)	0.998	-0.015 (-0.042 to 0.013)	0.845
Validation Cohort A	F1-score	-0.015 (-0.033 to 0.003)	0.941	0.052 (0.019 to 0.080)	0.003
Validation Cohort B	AUPRC	0.077 (0.027 to 0.134)	0.001	-0.064 (-0.161 to 0.048)	0.880
Validation Cohort B	F1-score	0.086 (0.038to 0.136)	<0.001	0.074 (-0.023 to 0.170)	0.069
Postoperative respiratory failure					
Derivation Cohort	AUPRC	0.034 (-0.001 to 0.069)	0.028	0.084 (0.049 to 0.127)	<0.001
Derivation Cohort	F1-score	0.055 (0.008 to 0.102)	0.007	0.147 (0.091 to 0.206)	<0.001
Validation Cohort A	AUPRC	0.033 (-0.001 to 0.067)	0.031	-0.039 (-0.093 to 0.015)	0.917
Validation Cohort A	F1-score	0.002 (-0.036 to 0.039)	0.457	0.006 (-0.032 to 0.047)	0.397
Validation Cohort B	AUPRC	0.014 (-0.121 to 0.135)	0.404	-0.154 (-0.315 to 0.045)	0.952
Validation Cohort B	F1-score	0.000 (-0.118 to 0.115)	0.485	-0.015 (-0.090 to 0.102)	0.670
In-hospital mortality					
Derivation Cohort	AUPRC	0.016 (-0.012 to 0.049)	0.141	0.051 (0.018 to 0.094)	<0.001
Derivation Cohort	F1-score	0.026 (-0.041 to 0.095)	0.239	0.094 (0.020 to 0.172)	0.003
Validation Cohort A	AUPRC	0.043 (0.011 to 0.078)	0.002	-0.028 (-0.093 to 0.040)	0.800
Validation Cohort A	F1-score	0.043 (0.002 to 0.083)	0.022	0.000 (-0.063 to 0.055)	0.487
Validation Cohort B	AUPRC	0.009 (-0.043 to 0.052)	0.319	-0.034 (-0.121 to 0.056)	0.779
Validation Cohort B	F1-score	0.021 (-0.028 to 0.070)	0.210	0.092 (0.032 to 0.156)	0.005

Comparisons were performed using bootstrapping paired t-tests with 1,000 bootstrap samples. Mean differences are shown with 95% confidence intervals and corresponding

p-values. MT-GBM: multitask gradient boosting machine, ASA: American Society of Anesthesiologists physical status classification, AUPRC: area under the precision-recall curve.

Supplementary Table S3. Comparisons of the area under the receiver operating characteristic curves for the multitask gradient boosting machine (MT-GBM) model, single-task prediction models with other tree-based algorithms, including Random Forest and XGBoost.

Cohorts	MT-GBM	Random Forest	<i>p</i> -value*	XGBoost	<i>p</i> -value*
	AUROC (95% CI)	AUROC (95% CI)		AUROC (95% CI)	
Acute kidney injury					
Derivation Cohort	0.805 (0.798-0.912)	0.796 (0.789–0.802)	<0.001	0.802 (0.795–0.809)	0.151
Validation Cohort A	0.789 (0.782–0.796)	0.766 (0.759–0.773)	<0.001	0.745 (0.738–0.753)	<0.001
Validation Cohort B	0.863 (0.850–0.876)	0.817 (0.802–0.831)	<0.001	0.807 (0.792–0.821)	<0.001
Postoperative respiratory failure					
Derivation Cohort	0.886 (0.880–0.891)	0.869 (0.863–0.874)	<0.001	0.878 (0.872–0.883)	0.010
Validation Cohort A	0.925 (0.920–0.929)	0.906 (0.901–0.911)	<0.001	0.886 (0.880–0.891)	<0.001
Validation Cohort B	0.911 (0.900–0.921)	0.907 (0.896–0.918)	0.246	0.888 (0.876–0.899)	0.001
In-hospital mortality					
Derivation Cohort	0.907 (0.902–0.912)	0.842 (0.835–0.848)	<0.001	0.865 (0.859–0.871)	<0.001
Validation Cohort A	0.913 (0.909–0.918)	0.889 (0.884–0.894)	<0.001	0.884 (0.878–0.889)	<0.001
Validation Cohort B	0.849 (0.835–0.862)	0.781 (0.766–0.797)	<0.001	0.804 (0.789–0.818)	<0.001

Comparisons were performed using DeLong's test. AUROC: area under the receiver operating characteristic curves, CI: confidence interval, MT-GBM: multitask gradient boosting machine, XGBoost: eXtreme Gradient Boosting.

Supplementary Table S4. Calibration metrics for the multitask gradient boosting machine model after spline recalibration across cohorts and outcomes

Cohorts	Calibration slope	Calibration intercept	Unreliability index P-value	Brier score
Acute kidney injury				
Derivation Cohort	0.634	0.012	<0.001	0.028
Validation Cohort A	0.477	0.023	<0.001	0.038
Validation Cohort B	0.888	0.004	0.215	0.030
Postoperative respiratory failure				
Derivation Cohort	0.502	0.005	<0.001	0.009
Validation Cohort A	0.739	0.007	<0.001	0.015
Validation Cohort B	0.520	0.005	<0.001	0.013
In-hospital mortality				
Derivation Cohort	0.555	0.003	0.050	0.005
Validation Cohort A	0.978	0.007	<0.001	0.013
Validation Cohort B	0.994	0.018	<0.001	0.027

Supplementary Table S5. Definitions of outliers in laboratory test results. The definition of outliers in laboratory findings was based on the methodology described in the MIMIC-Extract pipelines.

Variables	Valid low	Valid high
Hemoglobin, g/dL	0	25.0
White blood cell counts, 10 ³ /μL	0	1000
Platelet, 10 ³ /μL	0	2000
Blood urea nitrogen, mg/dL	0	250
Serum creatinine, mg/dL	0.1	60
Serum glucose, mg/dL	33	2000
Albumin, mg/dL	0.6	6.0
AST, IU/L	6	20,000
ALT, IU/L	2	10,000
Sodium, mmol/L	50	250
Potassium, mmol/L	0	12
PT INR*	0	10

AST: aspartate aminotransferase, ALT: alanine aminotransferase, PT: prothrombin time, INR: international normalized ratio, MIMIC: Medical Information Mart for Intensive Care. *: In the original MIMIC-Extract pipelines, the outlier value of prothrombin time was provided only in seconds; thus, we defined our valid range (0–10) for PT INR in this study.