

Supplement 1

eFigures & eTables

Clinicians vs. Artificial Intelligence in Patient Outcome Prediction in the
Intensive Care Unit

Kuang et al.

Table of Contents

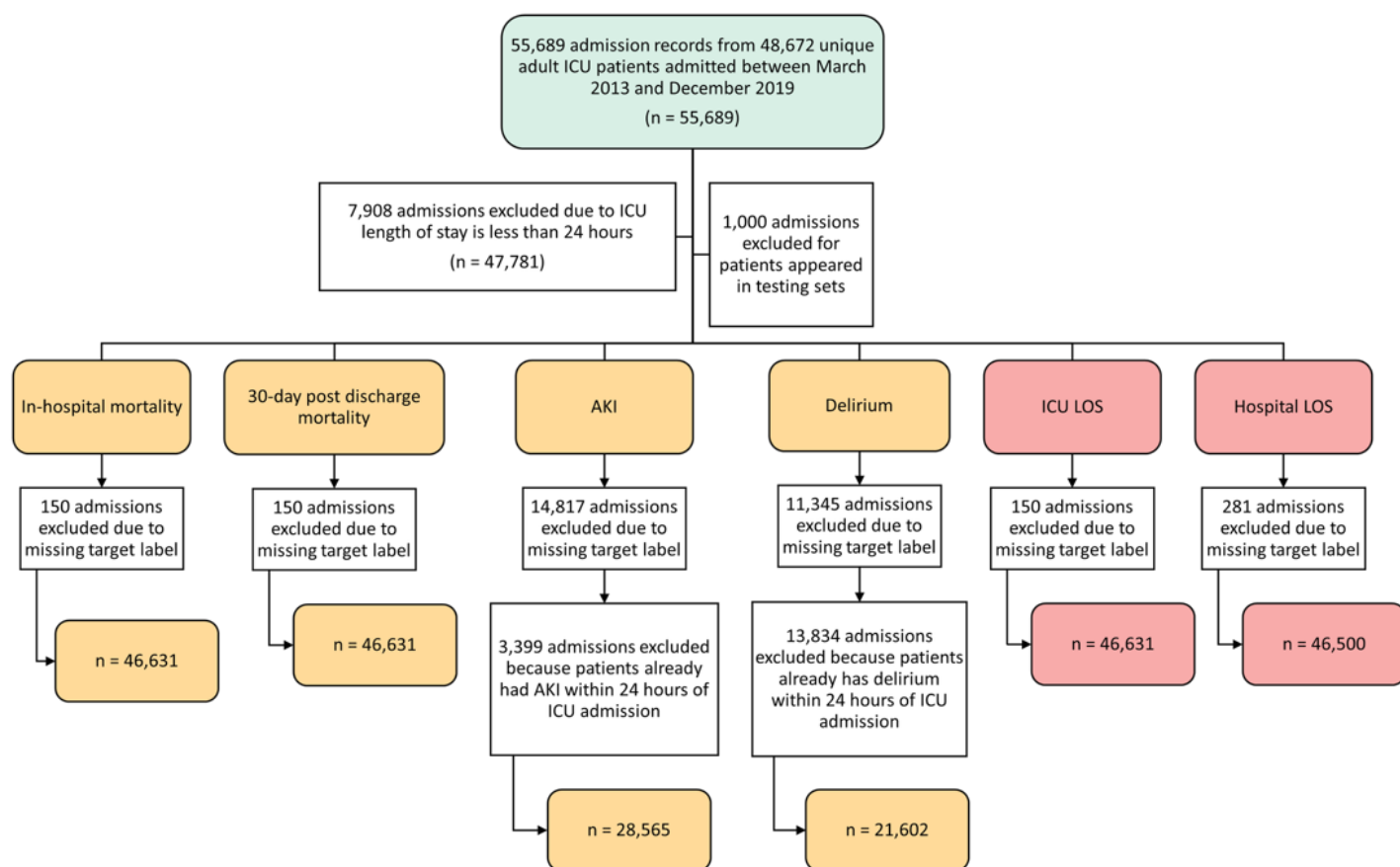
eFigure 1. AI model development cohort diagram for each predicted patient outcome.	4
eFigure 2. Machine learning pipeline for AI model training.	5
eFigure 3. Prospective setting patient cohort diagram for each predicted outcome.	6
eTable 1. Hyperparameter space for each classification model.	7
eTable 2. Hyperparameter space for each regression model.	8
eTable 3. Number of patient cases used in the retrospective setting stratified by the number of clinicians providing predictions.	9
eTable 4. Numbers of predictions made by different clinician types in the prospective setting.	10
eTable 5. Patient characteristics of the three cohorts used in this study.	11
eTable 6. Features selected for the AI models.	12
eTable 7. AI model performance from 5-fold cross-validation for classification outcomes.	18
eTable 8. AI model performance from 5-fold cross-validation for regression outcomes.	20
eTable 9. Retrospective comparison of clinician and AI prediction performances for the classification outcomes.	21
eTable 10. Retrospective comparison of clinician and AI prediction performances for the regression outcomes.	25
eTable 11. Prospective comparison of clinician and AI prediction performance for the classification outcomes.	26
eTable 12. Prospective comparison of clinician and AI prediction performance for the regression outcomes.	30
eTable 13. Inter-rater agreement between clinicians for the classification outcomes in the retrospective setting.	31
eTable 14. Inter-rater agreement between clinicians for the regression outcomes in the retrospective setting.	32
eTable 15. Inter-rater agreement between clinicians for the classification outcomes in the prospective setting.	33
eTable 16. Inter-rater agreement between clinicians for the regression outcomes in the prospective setting.	34
eTable 17. Inter-rater agreement between clinicians and AI for the classification outcomes in the retrospective setting.	35
eTable 18. Inter-rater agreement between clinicians and AI for the regression outcomes in the retrospective setting.	36

eTable 19. Inter-rater agreement between clinicians and AI for the classification outcomes in the prospective setting. 37

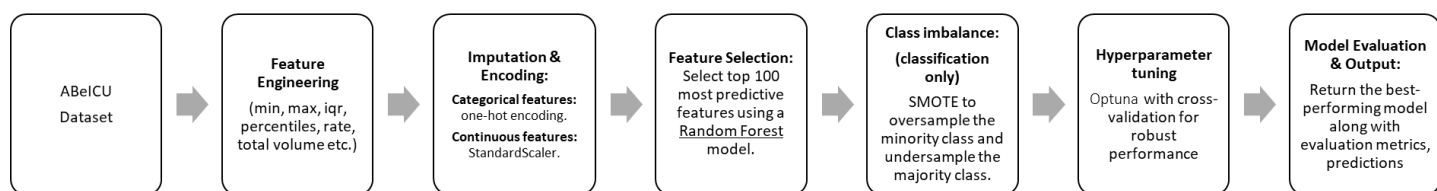
eTable 20. Inter-rater agreement between clinicians and AI for the regression outcomes in the prospective setting. 38

eTable 21. Likert scale threshold sensitivity analysis results for the retrospective comparison of clinician and AI prediction performances for the classification outcomes. 39

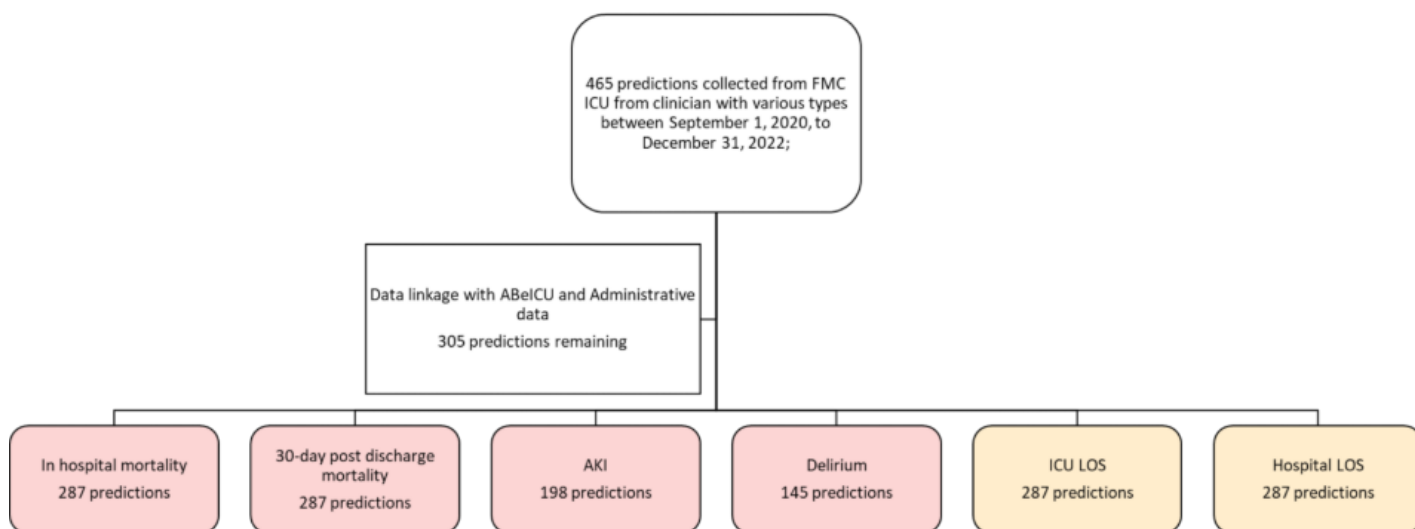
eTable 22. Likert scale threshold sensitivity analysis results for the prospective comparison of clinician and AI prediction performances for the classification outcomes. 43



eFigure 1. AI model development cohort diagram for each predicted patient outcome. AKI: acute kidney injury, FMC: Foothills Medical Centre, ICU: intensive care unit, LOS: length of stay.



eFigure 2. Machine learning pipeline for AI model training.



eFigure 3. Prospective setting patient cohort diagram for each predicted outcome. Collectively, participating clinicians made a total of 465 predictions for 287 patients. After linkage with administrative and ICU Tracer databases, 160 predictions were excluded due to identifier mismatch, leaving 305 predictions on 238 ICU admissions from 215 patients for inclusion in the study. AKI: acute kidney injury, FMC: Foothills Medical Centre, ICU: intensive care unit, LOS: length of stay.

eTable 1. Hyperparameter space for each classification model.

Model	Hyperparameter	Search Space
Logistic Regression	C	1×10^{-4} to 1×10^2 (log scale)
	Solver	lbfgs, newton-cg, liblinear, sag, saga
	Penalty	L2
	Max iterations	100 to 4000 (step 100)
Random Forest	n_estimators	100 to 3000 (step 50)
	max_depth	10 to 1500
	min_samples_split	2 to 20
	min_samples_leaf	1 to 20
	max_features	sqrt, log2, None
	criterion	entropy, gini
Support Vector Machine	C	1×10^{-5} to 10 (log scale)
	Kernel	linear, rbf, sigmoid
	Gamma	scale, auto
	class_weight	None, balanced
XGBoost	n_estimators	100 to 3000 (step 50)
	learning_rate	1×10^{-4} to 10 (log scale)
	max_depth	1 to 50
	subsample	0.5 to 1.0
	colsample_bytree	0.5 to 1.0
	gamma	0 to 5
	lambda	1×10^{-4} to 10 (log scale)
Neural Network	num_layers	1 to 5
	neurons_per_layer	16 to 512 (step 16), per layer
	dropout_rate	0.0 to 0.5
	learning_rate	1×10^{-4} to 1×10^{-1} (log scale)
	batch_size	{16, 32, 64, 128}
	epochs	10 to 100
	activation	relu, tanh, sigmoid
	optimizer	adam, rmsprop, sgd

eTable 2. Hyperparameter space for each regression model.

Model	Hyperparameter	Search Space
Elastic Net	alpha	1×10^{-5} to 10 (log scale)
	l1_ratio	0 to 1
Support Vector Machine	C	1×10^{-5} to 10 (log scale)
	kernel	linear, rbf, sigmoid
	gamma	1×10^{-5} to 10 (log scale)
XGBoost	n_estimators	100 to 3000 (step 50)
	learning_rate	1×10^{-4} to 1 (log scale)
	max_depth	3 to 15
	subsample	0.5 to 1.0
	colsample_bytree	0.5 to 1.0
	gamma	0 to 5
	lambda	1×10^{-4} to 10 (log scale)
Random Forest	n_estimators	100 to 1000 (step 100)
	max_depth	5 to 50
	min_samples_split	2 to 20
	min_samples_leaf	1 to 10
	max_features	sqrt, log2
	criterion	squared_error, friedman_mse
Neural Network	num_layers	1 to 5
	neurons_per_layer	16 to 512 (step 16), per layer (5 layers sampled)
	dropout_rate	0.0 to 0.5
	learning_rate	1×10^{-4} to 1×10^{-1} (log scale)
	batch_size	{16, 32, 64, 128}
	epochs	10 to 100
	activation	relu, tanh, sigmoid
	optimizer	adam, rmsprop, sgd

eTable 3. Number of patient cases used in the retrospective setting stratified by the number of clinicians providing predictions. Minor variations in sample size across outcomes are due to missing clinician predictions. AKI: acute kidney injury, ICU: intensive care unit.

Prediction Outcome	Group		
	1 Clinician	2 Clinicians	7 Clinicians
In-hospital mortality	575	374	50
30-day post-discharge mortality	574	376	50
AKI	570	380	50
Delirium	575	375	50
ICU length of stay	570	380	50
Hospital length of stay	571	379	50

eTable 4. Numbers of predictions made by different clinician types in the prospective setting.

Clinician Type	Clinician Subtype	Number of Predictions (%)	Grouped Number of Predictions (%)
Subspecialized physician	Attending physician	35 (9.8)	68 (19.0)
	Fellow	33 (9.2)	
Physician in training	Resident	72 (20.2)	86 (24.1)
	Intern	14 (3.9)	
Nurse	Registered nurse	203 (56.9)	203 (56.9)

eTable 5. Patient characteristics of the three cohorts used in this study. AKI: acute kidney injury, APACHE: Acute Physiology and Chronic Health Evaluation, ICU: intensive care unit, ICDSC: Intensive Care Delirium Screening Checklist, LOS: length of stay, SD: standard deviation, SOFA: Sequential Organ Failure Assessment.

Characteristic	Model Training Cohort	Retrospective Comparison Cohort	Prospective Comparison Cohort
Number of ICU Patients	41,096	990	215
Number of ICU Admissions	46,631	990	238
Period of ICU admission	February 2012 – December 2019	December 2018 – June 2019	September 2020 – December 2022
Male sex, n (%)	27781 (58.1)	588 (58.8)	132 (61.4)
Age in years, mean (SD)	56.9 (17.2)	58.32 (16.91)	53.8 (17.7)
Severity of illness scores at admission			
Glasgow Coma Scale	11.27	10.52	8.48
ICDSC – Delirium Score	2.24	2.19	2.55
Richmond Agitation Sedation Scale	–1.25	–1.58	–2.30
SOFA Score at Admission	6.61	6.5	7.66
Acute Physiological Score	15.55	14.76	16.99
APACHE II Score	19.98	19.33	20.6
APACHE III Score	67.7	63.88	70.17
Classification outcomes			
In-hospital mortality (%)	8,041 / 46,631 (17.24)	156 / 990 (15.76)	58 / 238 (34.4)
30-day post-discharge mortality (%)	8,254 / 46,631 (17.70)	160 / 990 (16.61)	63 / 238 (26.5)
AKI (%)	7,764 / 28,565 (27.18)	189 / 640 (29.53)	29 / 154 (18.8)
Delirium (%)	10,672 / 21,602 (49.40)	240 / 544 (44.12)	79 / 113 (69.9)
Regression outcomes			
ICU LOS in days, mean (SD)	6.7 (9.2)	7.7 (8.7)	7.2 (8.2)
Hospital LOS in days, mean (SD)	23.6 (39.1)	24.3 (27.8)	50.8 (73.8)

eTable 6. Features selected for the AI models. This list is the super set of all features used by all classification and regression models, each of which used 100 features.

Index	Feature
1	Gender
2	Eye Opening average value
3	Eye Opening median value
4	Eye Opening min value
5	Eye Opening percentile 25
6	Eye Opening percentile 75
7	Verbal Response median value
8	Verbal Response min value
9	Verbal Response percentile 25
10	Glasgow Coma Scale average value
11	Glasgow Coma Scale median value
12	Glasgow Coma Scale min value
13	Glasgow Coma Scale range
14	Motor Response average value
15	Motor Response median value
16	Motor Response percentile 25
17	Motor Response percentile 75
18	Intensive Care Delirium Screening Checklist average value
19	Intensive Care Delirium Screening Checklist range
20	Inattention average value
21	Inattention median value
22	Inattention range
23	Inattention percentile 25
24	Psychomotor Agitation or Retardation average value
25	Psychomotor Agitation or Retardation median value
26	Psychomotor Agitation or Retardation range
27	Psychomotor Agitation or Retardation percentile 25
28	Fluctuations average value
29	Fluctuations max value
30	Fluctuations range
31	Fluctuations percentile 25
32	Altered Level of Consciousness average value
33	Altered Level of Consciousness median value
34	Altered Level of Consciousness range
35	Altered Level of Consciousness percentile 25
36	Hallucination Delusion or Psychosis average value
37	Hallucination Delusion or Psychosis max value
38	Hallucination Delusion or Psychosis median value

39	Hallucination Delusion or Psychosis percentile 25
40	Hallucination Delusion or Psychosis percentile 75
41	Sleep Wake Cycle Disturbance average value
42	Sleep Wake Cycle Disturbance max value
43	Sleep Wake Cycle Disturbance median value
44	Sleep Wake Cycle Disturbance min value
45	Sleep Wake Cycle Disturbance standard deviation
46	Sleep Wake Cycle Disturbance percentile 25
47	Disorientation average value
48	Disorientation max value
49	Disorientation median value
50	Inappropriate Mood or Speech max value
51	Inappropriate Mood or Speech median value
52	Inappropriate Mood or Speech standard deviation
53	Inappropriate Mood or Speech percentile 25
54	Richmond Agitation Sedation Scale average value
55	Richmond Agitation Sedation Scale max value
56	Richmond Agitation Sedation Scale standard deviation
57	Urine Indwelling Catheter average value
58	Urine Indwelling Catheter standard deviation
59	Heart Rate average value
60	Heart Rate max value
61	Heart Rate median value
62	Heart Rate standard deviation
63	Oxygen Saturation average value
64	Oxygen Saturation max value
65	Oxygen Saturation standard deviation
66	Oxygen Saturation - Site 1 average value
67	Oxygen Saturation - Site 1 max value
68	Oxygen Saturation - Site 1 median value
69	Oxygen Saturation - Site 1 iqr
70	Oxygen Saturation - Site 1 percentile 25
71	Non-Invasive Blood Pressure - Systolic average value
72	Non-Invasive Blood Pressure - Systolic max value
73	Non-Invasive Blood Pressure - Systolic percentile 25
74	Non-Invasive Blood Pressure - Diastolic max value
75	Non-Invasive Blood Pressure - Diastolic median value
76	Non-Invasive Blood Pressure - Diastolic min value
77	Non-Invasive Blood Pressure - Mean average value
78	Non-Invasive Blood Pressure - Mean max value
79	Non-Invasive Blood Pressure - Mean median value

80	Non-Invasive Blood Pressure - Mean min value
81	Non-Invasive Blood Pressure - Mean iqr
82	Resp Rate Bedside Monitor average value
83	Resp Rate Bedside Monitor max value
84	Resp Rate Bedside Monitor percentile 25
85	Resp Rate Bedside Monitor iqr
86	Urea Blood average value
87	Urea Blood max value
88	Urea Blood iqr
89	Carbon Dioxide Content Blood average value
90	Carbon Dioxide Content Blood max value
91	Carbon Dioxide Content Blood median value
92	Carbon Dioxide Content Blood min value
93	Carbon Dioxide Content Blood iqr
94	Creatinine Blood average value
95	Creatinine Blood max value
96	Creatinine Blood median value
97	Creatinine Blood min value
98	Potassium Blood max value
99	Potassium Blood median value
100	Sodium Blood average value
101	Sodium Blood max value
102	Sodium Blood median value
103	Sodium Blood min value
104	Sodium Blood Gas average value
105	Sodium Blood Gas max value
106	Sodium Blood Gas median value
107	Sodium Blood Gas min value
108	Partial Pressure of Carbon Dioxide - Arterial max value
109	Partial Pressure of Carbon Dioxide - Arterial median value
110	Partial Pressure of Carbon Dioxide - Arterial min value
111	Potential of Hydrogen - Arterial average value
112	Potential of Hydrogen - Arterial max value
113	Potential of Hydrogen - Arterial median value
114	Potential of Hydrogen - Arterial range
115	Partial Pressure of Oxygen - Arterial average value
116	Partial Pressure of Oxygen - Arterial max value
117	Partial Pressure of Oxygen - Arterial median value
118	Partial Pressure of Oxygen - Arterial min value
119	Partial Pressure of Oxygen - Arterial percentile 25
120	Total Hemoglobin Blood Gas average value

121	Total Hemoglobin Blood Gas max value
122	Total Hemoglobin Blood Gas median value
123	Total Hemoglobin Blood Gas min value
124	Potassium Blood Gas average value
125	Potassium Blood Gas max value
126	Potassium Blood Gas median value
127	Potassium Blood Gas min value
128	Potassium Blood Gas percentile 25
129	Hematocrit average value
130	Hematocrit max value
131	Hematocrit median value
132	Hematocrit min value
133	Hemoglobin max value
134	Hemoglobin median value
135	Hemoglobin min value
136	Red Blood Cells average value
137	Red Blood Cells median value
138	Red Blood Cells percentile 25
139	White Blood Cells average value
140	White Blood Cells median value
141	White Blood Cells min value
142	White Blood Cells percentile 25
143	Respiratory (admission) average value
144	Respiratory (admission) median value
145	Respiratory (admission) min value
146	Respiratory (admission) standard deviation
147	Respiratory (admission) percentile 25
148	Coagulatory (admission) max value
149	Coagulatory (admission) median value
150	Coagulatory (admission) min value
151	Coagulatory (admission) range
152	Liver (admission) average value
153	Liver (admission) median value
154	Liver (admission) standard deviation
155	Liver (admission) percentile 75
156	Cardiovascular (admission) average value
157	Cardiovascular (admission) min value
158	Cardiovascular (admission) percentile 25
159	Cardiovascular (admission) standard deviation
160	Cardiovascular (admission) range
161	Neurologic (admission) average value

162	Neurologic (admission) max value
163	Neurologic (admission) median value
164	Neurologic (admission) min value
165	Neurologic (admission) standard deviation
166	Neurologic (admission) range
167	Renal (admission) average value
168	Renal (admission) max value
169	Renal (admission) median value
170	Renal (admission) standard deviation
171	Renal (admission) range
172	Renal (admission) percentile 25
173	Sequential Organ Failure Assessment (admission) max value
174	Sequential Organ Failure Assessment (admission) standard deviation
175	Sequential Organ Failure Assessment (admission) range
176	Age Point max value
177	Age Point standard deviation
178	Age Point range
179	Age Point percentile 25
180	Age Point percentile 75
181	Chronic Health Point percentile 25
182	Acute Physiological Score max value
183	Acute Physiological Score min value
184	Acute Physiological Score median value
185	Acute Physiological Score percentile 25
186	Apache II Score max value
187	Apache II Score median value
188	Apache II Score min value
189	Apache II Score range
190	Mortality Prediction Model median value
191	Mortality Prediction Model percentile 75
192	Apache II Compliance min value
193	Apache II Compliance percentile 25
194	Apache II Compliance range
195	Acute Physiological Score min value
196	Acute Physiological Score range
197	Acute Physiological Score percentile 75
198	Age Point average value
199	Age Point max value
200	Age Point median value
201	Age Point min value
202	Age Point percentile 25

203	Age Point percentile 75
204	Chronic Health Point max value
205	Chronic Health Point min value
206	Chronic Health Point percentile 25
207	Chronic Health Point percentile 75
208	Apache III Score median value
209	Apache III Score min value
210	Apache III Score percentile 25
211	Apache III Score percentile 75
212	Apache III Compliance median value
213	Apache III Compliance min value
214	Apache III Compliance percentile 25
215	Apache III Compliance percentile 75
216	0.9% NaCl average rate
217	0.9% NaCl iqr rate
218	0.9% NaCl median value rate
219	0.9% NaCl min rate
220	0.9% NaCl percentile 25 rate
221	0.9% NaCl percentile 75 rate
222	0.9% NaCl rate range
223	0.9% NaCl standard deviation rate
224	0.9% NaCl total time
225	0.9% NaCl total volume

eTable 7. AI model performance from 5-fold cross-validation for classification outcomes. Based on median AUROC, the best model for every outcome was XGBoost. AKI: acute kidney injury, AUROC: area under the receiver operating characteristic curve, AUPRC: area under the precision-recall curve, IQR: interquartile range.

Outcome	Model	Accuracy, median (IQR)	Balanced Accuracy, median (IQR)	Specificity, median (IQR)	Sensitivity, median (IQR)	F1, median (IQR)	AUROC, median (IQR)	AUPRC, median (IQR)
In-hospital mortality	Logistic regression	0.80 (0.78– 0.83)	0.67 (0.64– 0.70)	0.87 (0.84– 0.90)	0.48 (0.45– 0.51)	0.46 (0.43– 0.49)	0.85 (0.82– 0.88)	0.28 (0.25– 0.31)
	Support vector machine	0.81 (0.78– 0.84)	0.69 (0.66– 0.72)	0.86 (0.83– 0.89)	0.55 (0.52– 0.58)	0.49 (0.46– 0.52)	0.85 (0.81– 0.88)	0.30 (0.27– 0.33)
	Random forest	0.81 (0.79– 0.84)	0.68 (0.65– 0.71)	0.86 (0.83– 0.89)	0.52 (0.49– 0.55)	0.48 (0.45– 0.51)	0.86 (0.83– 0.89)	0.29 (0.26– 0.32)
	XGBoost	0.83 (0.80– 0.86)	0.69 (0.66– 0.72)	0.88 (0.85– 0.91)	0.52 (0.49– 0.55)	0.50 (0.47– 0.53)	0.87 (0.84– 0.90)	0.31 (0.28– 0.34)
	Neural network	0.84 (0.81– 0.87)	0.66 (0.63– 0.69)	0.92 (0.89– 0.95)	0.42 (0.39– 0.45)	0.47 (0.44– 0.50)	0.86 (0.82– 0.89)	0.31 (0.28– 0.34)
30-day post- discharge mortality	Logistic regression	0.82 (0.79– 0.85)	0.66 (0.63– 0.69)	0.88 (0.85– 0.91)	0.47 (0.44– 0.50)	0.46 (0.43– 0.49)	0.85 (0.81– 0.88)	0.29 (0.26– 0.32)
	Support vector machine	0.81 (0.78– 0.84)	0.67 (0.64– 0.70)	0.87 (0.84– 0.90)	0.49 (0.46– 0.52)	0.47 (0.44– 0.50)	0.85 (0.82– 0.89)	0.29 (0.26– 0.32)
	Random forest	0.81 (0.78– 0.84)	0.68 (0.65– 0.71)	0.84 (0.81– 0.87)	0.49 (0.46– 0.52)	0.47 (0.44– 0.50)	0.86 (0.82– 0.90)	0.30 (0.27– 0.33)
	XGBoost	0.84 (0.81– 0.87)	0.67 (0.64– 0.70)	0.92 (0.89– 0.95)	0.44 (0.41– 0.47)	0.49 (0.46– 0.52)	0.87 (0.83– 0.91)	0.32 (0.29– 0.35)

	Neural network	0.84 (0.81–0.87)	0.65 (0.62–0.68)	0.93 (0.90–0.96)	0.39 (0.36–0.42)	0.46 (0.43–0.49)	0.86 (0.82–0.90)	0.30 (0.27–0.33)
AKI	Logistic regression	0.73 (0.70–0.76)	0.63 (0.60–0.66)	0.86 (0.83–0.89)	0.41 (0.38–0.44)	0.46 (0.43–0.49)	0.71 (0.68–0.74)	0.40 (0.37–0.43)
	Support vector machine	0.73 (0.70–0.76)	0.62 (0.59–0.65)	0.85 (0.82–0.88)	0.40 (0.37–0.43)	0.45 (0.42–0.48)	0.74 (0.70–0.78)	0.39 (0.36–0.42)
	Random forest	0.75 (0.72–0.78)	0.65 (0.62–0.68)	0.88 (0.85–0.91)	0.45 (0.42–0.48)	0.50 (0.47–0.53)	0.75 (0.71–0.79)	0.42 (0.39–0.45)
	XGBoost	0.74 (0.71–0.77)	0.67 (0.64–0.70)	0.83 (0.80–0.86)	0.52 (0.49–0.55)	0.52 (0.49–0.55)	0.76 (0.72–0.80)	0.42 (0.39–0.45)
	Neural network	0.73 (0.70–0.76)	0.65 (0.62–0.68)	0.84 (0.81–0.87)	0.48 (0.45–0.51)	0.50 (0.47–0.53)	0.74 (0.70–0.78)	0.41 (0.38–0.44)
Delirium	Logistic regression	0.74 (0.71–0.77)	0.74 (0.71–0.77)	0.71 (0.68–0.74)	0.76 (0.73–0.79)	0.76 (0.73–0.79)	0.83 (0.79–0.86)	0.71 (0.68–0.74)
	Support vector machine	0.77 (0.74–0.80)	0.77 (0.74–0.80)	0.75 (0.72–0.78)	0.79 (0.76–0.82)	0.78 (0.75–0.81)	0.83 (0.80–0.87)	0.74 (0.71–0.77)
	Random forest	0.74 (0.71–0.77)	0.75 (0.72–0.78)	0.76 (0.73–0.79)	0.74 (0.71–0.77)	0.75 (0.72–0.78)	0.84 (0.80–0.88)	0.72 (0.69–0.75)
	XGBoost	0.76 (0.73–0.79)	0.76 (0.73–0.79)	0.76 (0.73–0.79)	0.75 (0.72–0.78)	0.77 (0.74–0.80)	0.85 (0.81–0.89)	0.73 (0.70–0.76)
	Neural network	0.74 (0.71–0.77)	0.74 (0.71–0.77)	0.78 (0.75–0.81)	0.70 (0.67–0.73)	0.74 (0.71–0.77)	0.83 (0.80–0.86)	0.72 (0.69–0.75)

eTable 8. AI model performance from 5-fold cross-validation for regression outcomes. Based on median RMSE, the best model for every outcome was XGBoost. AKI: acute kidney injury, AUROC: area under the receiver operating characteristic curve, AUPRC: area under the precision-recall curve, ICU: intensive care unit, IQR: interquartile range, LOS: length of stay, MAE: mean absolute error, RMSE: root mean squared error.

Outcome	Model	MAE in days, median (IQR)	RMSE in days, median (IQR)	R ² , median (IQR)
ICU LOS	Elastic net	5.40 (4.39-5.16)	8.58 (8.07-9.09)	0.28 (0.23-0.33)
	Support vector machine	5.15 (4.14-4.99)	8.62 (8.12-9.17)	0.26 (0.20-0.31)
	Random forest	5.32 (4.30-5.10)	8.45 (7.97-8.99)	0.31 (0.26-0.36)
	XGBoost	5.30 (4.34-5.11)	8.40 (7.92-8.91)	0.25 (0.17-0.30)
	Neural network	5.38 (4.37-5.17)	8.46 (7.99-9.01)	0.30 (0.24-0.36)
Hospital LOS	Elastic net	18.36 (17.17-19.57)	38.31 (36.34-40.56)	0.22 (0.16-0.27)
	Support vector machine	15.64 (14.25-17.14)	39.50 (37.36-42.01)	0.16 (0.11-0.22)
	Random forest	18.00 (16.78-19.25)	37.85 (35.78-40.07)	0.24 (0.18-0.29)
	XGBoost	17.99 (16.86-19.29)	37.42 (35.43-39.76)	0.21 (0.15-0.24)
	Neural network	17.90 (16.68-19.19)	37.56 (35.57-39.95)	0.24 (0.18-0.30)

eTable 9. Retrospective comparison of clinician and AI prediction performances for the classification outcomes. Results are stratified by the number of clinicians providing predictions: one, two, and seven clinicians. Statistical significance was assessed using two-sample Wilcoxon rank-sum tests. AKI: acute kidney injury, CI: confidence interval, NPV: negative predictive value, PPV: positive predictive value.

Outcome	Metric	Group	Clinicians, median (95% CI)	AI, median (95% CI)	P Value
In-hospital Mortality	Accuracy	One clinician	0.806 (0.774–0.838)	0.864 (0.836–0.891)	<.001
		Two clinicians	0.828 (0.790–0.863)	0.836 (0.798–0.871)	<.001
		Seven clinicians	0.800 (0.680–0.900)	0.760 (0.640–0.880)	<.001
	Balanced Accuracy	One clinician	0.656 (0.601–0.707)	0.732 (0.675–0.785)	<.001
		Two clinicians	0.609 (0.553–0.670)	0.704 (0.641–0.777)	<.001
		Seven clinicians	0.628 (0.436–0.811)	0.538 (0.409–0.744)	<.001
	PPV	One clinician	0.391 (0.290–0.485)	0.575 (0.459–0.673)	<.001
		Two clinicians	0.421 (0.278–0.576)	0.475 (0.355–0.593)	<.001
		Seven clinicians	0.375 (0.000–0.750)	0.231 (0.000–0.571)	<.001
	Specificity	One clinician	0.874 (0.847–0.902)	0.925 (0.899–0.947)	<.001
		Two clinicians	0.925 (0.895–0.953)	0.896 (0.861–0.926)	<.001
		Seven clinicians	0.884 (0.778–0.976)	0.860 (0.756–0.952)	<.001
	Sensitivity	One clinician	0.436 (0.337–0.538)	0.538 (0.432–0.647)	<.001
		Two clinicians	0.293 (0.183–0.414)	0.514 (0.386–0.655)	<.001
		Seven clinicians	0.375 (0.000–0.715)	0.222 (0.000–0.600)	<.001
	F1 Score	One clinician	0.413 (0.319–0.495)	0.555 (0.460–0.639)	<.001
		Two clinicians	0.347 (0.226–0.463)	0.493 (0.379–0.602)	<.001
		Seven clinicians	0.364 (0.000–0.637)	0.222 (0.000–0.526)	<.001
	NPV	One clinician	0.894 (0.865–0.923)	0.915 (0.889–0.939)	<.001
		Two clinicians	0.877 (0.841–0.909)	0.910 (0.876–0.940)	<.001

		Seven clinicians	0.884 (0.775–0.975)	0.857 (0.750–0.954)	<.001
30-Day Post-Discharge Mortality	Accuracy	One clinician	0.807 (0.774–0.837)	0.869 (0.839–0.896)	<.001
		Two clinicians	0.821 (0.783–0.861)	0.861 (0.823–0.893)	<.001
		Seven clinicians	0.860 (0.760–0.940)	0.800 (0.680–0.900)	<.001
	Balanced Accuracy	One clinician	0.620 (0.570–0.673)	0.694 (0.639–0.748)	<.001
		Two clinicians	0.550 (0.506–0.601)	0.690 (0.627–0.753)	<.001
		Seven clinicians	0.725 (0.560–0.890)	0.565 (0.464–0.718)	<.001
	PPV	One clinician	0.373 (0.262–0.478)	0.618 (0.500–0.745)	<.001
		Two clinicians	0.375 (0.182–0.588)	0.590 (0.439–0.737)	<.001
		Seven clinicians	0.727 (0.333–1.000)	0.500 (0.000–1.000)	<.001
	Specificity	One clinician	0.892 (0.863–0.918)	0.949 (0.929–0.970)	<.001
		Two clinicians	0.953 (0.929–0.974)	0.941 (0.912–0.965)	<.001
		Seven clinicians	0.952 (0.872–1.000)	0.951 (0.878–1.000)	0.65
	Sensitivity	One clinician	0.347 (0.250–0.452)	0.440 (0.330–0.547)	<.001
		Two clinicians	0.145 (0.064–0.246)	0.439 (0.315–0.564)	<.001
		Seven clinicians	0.500 (0.167–0.833)	0.182 (0.000–0.500)	<.001
	F1 Score	One clinician	0.362 (0.266–0.452)	0.513 (0.407–0.604)	<.001
		Two clinicians	0.209 (0.095–0.337)	0.500 (0.379–0.615)	<.001
		Seven clinicians	0.588 (0.222–0.833)	0.267 (0.000–0.571)	<.001
	NPV	One clinician	0.881 (0.851–0.909)	0.900 (0.875–0.925)	<.001
		Two clinicians	0.853 (0.816–0.887)	0.897 (0.862–0.927)	<.001
		Seven clinicians	0.886 (0.786–0.976)	0.830 (0.711–0.918)	<.001
AKI	Accuracy	One clinician	0.691 (0.645–0.741)	0.768 (0.719–0.809)	<.001

		Two clinicians	0.622 (0.560–0.680)	0.747 (0.693–0.801)	<.001
		Seven clinicians	0.697 (0.545–0.848)	0.758 (0.606–0.879)	<.001
	Balanced Accuracy	One clinician	0.679 (0.628–0.736)	0.708 (0.653–0.761)	<.001
		Two clinicians	0.599 (0.530–0.661)	0.667 (0.602–0.735)	<.001
		Seven clinicians	0.704 (0.500–0.889)	0.523 (0.417–0.682)	<.001
	PPV	One clinician	0.497 (0.414–0.578)	0.637 (0.535–0.731)	<.001
		Two clinicians	0.392 (0.296–0.480)	0.585 (0.440–0.705)	<.001
		Seven clinicians	0.375 (0.125–0.643)	0.333 (0.000–1.000)	<.001
	Specificity	One clinician	0.709 (0.652–0.764)	0.860 (0.816–0.903)	<.001
		Two clinicians	0.655 (0.588–0.728)	0.863 (0.807–0.911)	<.001
		Seven clinicians	0.700 (0.520–0.852)	0.926 (0.808–1.000)	<.001
	Sensitivity	One clinician	0.652 (0.564–0.738)	0.551 (0.464–0.650)	<.001
		Two clinicians	0.543 (0.434–0.652)	0.473 (0.358–0.592)	<.001
		Seven clinicians	0.714 (0.333–1.000)	0.125 (0.000–0.429)	<.001
	F1 Score	One clinician	0.563 (0.487–0.638)	0.591 (0.510–0.670)	<.001
		Two clinicians	0.454 (0.355–0.532)	0.520 (0.404–0.624)	<.001
		Seven clinicians	0.500 (0.182–0.737)	0.182 (0.000–0.500)	<.001
	NPV	One clinician	0.822 (0.766–0.870)	0.815 (0.767–0.860)	<.001
		Two clinicians	0.779 (0.705–0.844)	0.798 (0.740–0.856)	<.001
		Seven clinicians	0.900 (0.762–1.000)	0.800 (0.636–0.931)	<.001
Delirium	Accuracy	One clinician	0.616 (0.559–0.667)	0.800 (0.752–0.838)	<.001
		Two clinicians	0.618 (0.549–0.681)	0.784 (0.730–0.838)	<.001
		Seven clinicians	0.680 (0.480–0.840)	0.760 (0.599–0.920)	<.001

	Balanced Accuracy	One clinician	0.617 (0.560–0.667)	0.798 (0.751–0.837)	<.001
		Two clinicians	0.623 (0.555–0.686)	0.778 (0.721–0.837)	<.001
		Seven clinicians	0.683 (0.497–0.850)	0.763 (0.566–0.920)	<.001
	PPV	One clinician	0.658 (0.576–0.729)	0.810 (0.750–0.861)	<.001
		Two clinicians	0.721 (0.631–0.802)	0.830 (0.761–0.896)	<.001
		Seven clinicians	0.714 (0.375–1.000)	0.750 (0.462–1.000)	<.001
	Specificity	One clinician	0.628 (0.547–0.701)	0.775 (0.704–0.836)	<.001
		Two clinicians	0.659 (0.554–0.753)	0.753 (0.654–0.845)	<.001
		Seven clinicians	0.786 (0.533–1.000)	0.769 (0.500–1.000)	0.007
	Sensitivity	One clinician	0.605 (0.533–0.677)	0.823 (0.764–0.876)	<.001
		Two clinicians	0.587 (0.500–0.672)	0.805 (0.733–0.868)	<.001
		Seven clinicians	0.583 (0.300–0.857)	0.769 (0.500–1.000)	<.001
	F1 Score	One clinician	0.630 (0.562–0.692)	0.815 (0.771–0.857)	<.001
		Two clinicians	0.646 (0.567–0.716)	0.817 (0.762–0.865)	<.001
		Seven clinicians	0.636 (0.363–0.833)	0.750 (0.500–0.917)	<.001
	NPV	One clinician	0.573 (0.506–0.653)	0.789 (0.724–0.850)	<.001
		Two clinicians	0.515 (0.419–0.615)	0.720 (0.624–0.813)	<.001
		Seven clinicians	0.667 (0.438–0.900)	0.786 (0.545–1.000)	<.001

eTable10. Retrospective comparison of clinician and AI prediction performances for the regression outcomes. Results are stratified by the number of clinicians providing predictions: one, two, and seven clinicians. Statistical significance was assessed using two-sample Wilcoxon rank-sum tests. CI: confidence interval, LOS: length of stay, MAE: mean absolute error, RMSE: root mean squared error.

Outcome	Metric	Group	Clinicians, median (95% CI)	AI, median (95% CI)	P value
ICU LOS	MAE	One clinician	6.144 (5.837–6.504)	5.126 (4.841–5.466)	<.001
		Two clinicians	4.436 (4.109–4.786)	4.534 (4.275–4.800)	<.001
		Seven clinicians	7.641 (6.384–9.123)	7.645 (6.374–9.019)	0.533
	RMSE	One clinician	10.393 (9.676–11.196)	9.394 (8.416–10.348)	<.001
		Two clinicians	8.325 (7.538–9.176)	7.227 (6.557–7.889)	<.001
		Seven clinicians	13.600 (10.678–16.753)	13.272 (10.088–16.704)	<.001
	R ²	One clinician	–0.094 (–0.205 to –0.012)	0.110 (0.083–0.142)	<.001
		Two clinicians	–0.064 (–0.131 to –0.010)	0.201 (0.170–0.228)	<.001
		Seven clinicians	–0.077 (–0.146 to –0.025)	–0.045 (–0.111 to 0.054)	<.001
Hospital LOS	MAE	One clinician	18.70 (17.56–19.80)	18.53 (17.70–19.42)	<.001
		Two clinicians	16.73 (15.46–18.08)	17.14 (16.18–18.08)	<.001
		Seven clinicians	34.23 (27.84–41.58)	33.53 (28.40–39.82)	<.001
	RMSE	One clinician	31.96 (29.12–34.78)	26.96 (25.27–28.58)	<.001
		Two clinicians	29.12 (26.68–31.37)	24.84 (22.84–26.64)	<.001
		Seven clinicians	59.13 (48.32–69.97)	53.33 (43.77–62.88)	<.001
	R ²	One clinician	–0.284 (–0.526 to –0.119)	0.082 (0.047–0.114)	<.001
		Two clinicians	–0.190 (–0.236 to –0.141)	0.132 (0.085–0.182)	<.001
		Seven clinicians	–0.271 (–0.371 to –0.190)	–0.031 (–0.101 to 0.028)	<.001

eTable 11. Prospective comparison of clinician and AI prediction performance for the classification outcomes. Results are stratified by clinician type: subspecialized physicians, physicians in training, and nurses. Statistical significance was assessed using two-sample Wilcoxon rank-sum tests. AKI: acute kidney injury, CI: confidence interval, NPV: negative predictive value, PPV: positive predictive value.

Outcome	Metric	Group	Clinician, median (95% CI)	AI, median (95% CI)	P Value
In-Hospital Mortality	Accuracy	Subspecialized physician	0.852 (0.754–0.934)	0.820 (0.705–0.902)	<.001
		Physician in training	0.800 (0.700–0.886)	0.743 (0.643–0.843)	<.001
		Nurse	0.801 (0.744–0.865)	0.814 (0.756–0.872)	<.001
	Balanced Accuracy	Subspecialized physician	0.781 (0.652–0.897)	0.782 (0.636–0.902)	0.924
		Physician in training	0.676 (0.563–0.792)	0.687 (0.567–0.810)	<.001
		Nurse	0.665 (0.583–0.757)	0.694 (0.606–0.780)	<.001
	PPV	Subspecialized physician	0.778 (0.500–1.000)	0.649 (0.375–0.867)	<.001
		Physician in training	0.812 (0.500–1.000)	0.550 (0.333–0.750)	<.001
		Nurse	0.609 (0.400–0.786)	0.632 (0.429–0.808)	<.001
	Specificity	Subspecialized physician	0.935 (0.844–1.000)	0.867 (0.755–0.955)	<.001
		Physician in training	0.962 (0.900–1.000)	0.820 (0.712–0.918)	<.001
		Nurse	0.918 (0.868–0.966)	0.918 (0.867–0.963)	0.468
	Sensitivity	Subspecialized physician	0.625 (0.384–0.857)	0.700 (0.444–0.923)	<.001
		Physician in training	0.391 (0.176–0.625)	0.556 (0.333–0.783)	<.001
		Nurse	0.414 (0.256–0.581)	0.471 (0.314–0.629)	<.001
	F1 Score	Subspecialized physician	0.692 (0.462–0.857)	0.667 (0.435–0.833)	<.001
		Physician in training	0.528 (0.267–0.727)	0.550 (0.348–0.727)	<.001
		Nurse	0.491 (0.326–0.640)	0.540 (0.381–0.667)	<.001
	NPV	Subspecialized physician	0.880 (0.778–0.959)	0.891 (0.786–0.977)	<.001
		Physician in training	0.803 (0.695–0.887)	0.824 (0.714–0.923)	<.001
		Nurse	0.840 (0.776–0.901)	0.854 (0.791–0.909)	<.001

30-Day Post-Discharge Mortality	Accuracy	Subspecialized physician	0.885 (0.803–0.951)	0.836 (0.738–0.918)	<.001
		Physician in training	0.829 (0.729–0.900)	0.757 (0.643–0.843)	<.001
		Nurse	0.821 (0.756–0.878)	0.808 (0.744–0.865)	<.001
	Balanced Accuracy	Subspecialized physician	0.814 (0.692–0.922)	0.782 (0.647–0.900)	<.001
		Physician in training	0.711 (0.599–0.825)	0.680 (0.560–0.795)	<.001
		Nurse	0.692 (0.609–0.785)	0.665 (0.587–0.751)	<.001
	PPV	Subspecialized physician	0.923 (0.714–1.000)	0.733 (0.467–0.938)	<.001
		Physician in training	0.900 (0.667–1.000)	0.586 (0.353–0.812)	<.001
		Nurse	0.719 (0.520–0.875)	0.682 (0.476–0.875)	<.001
	Specificity	Subspecialized physician	0.978 (0.929–1.000)	0.909 (0.814–0.978)	<.001
		Physician in training	0.980 (0.936–1.000)	0.860 (0.759–0.944)	<.001
		Nurse	0.943 (0.897–0.976)	0.941 (0.897–0.982)	0.189
	Sensitivity	Subspecialized physician	0.652 (0.400–0.857)	0.652 (0.412–0.875)	0.755
		Physician in training	0.439 (0.217–0.667)	0.500 (0.292–0.706)	<.001
		Nurse	0.442 (0.290–0.622)	0.391 (0.233–0.556)	<.001
	F1 Score	Subspecialized physician	0.759 (0.545–0.905)	0.690 (0.467–0.849)	<.001
		Physician in training	0.593 (0.333–0.786)	0.538 (0.320–0.703)	<.001
		Nurse	0.548 (0.379–0.693)	0.493 (0.327–0.646)	<.001
	NPV	Subspecialized physician	0.880 (0.784–0.960)	0.872 (0.762–0.957)	0.001
		Physician in training	0.817 (0.712–0.902)	0.815 (0.702–0.906)	0.18
		Nurse	0.841 (0.773–0.903)	0.828 (0.759–0.890)	<.001
AKI	Accuracy	Subspecialized physician	0.844 (0.733–0.956)	0.800 (0.689–0.911)	<.001
		Physician in training	0.580 (0.440–0.720)	0.760 (0.640–0.880)	<.001
		Nurse	0.699 (0.612–0.786)	0.796 (0.718–0.874)	<.001

	Balanced Accuracy	Subspecialized physician	0.805 (0.648–0.951)	0.717 (0.555–0.867)	<.001
		Physician in training	0.583 (0.401–0.758)	0.621 (0.455–0.804)	<.001
		Nurse	0.605 (0.485–0.733)	0.645 (0.531–0.753)	<.001
	PPV	Subspecialized physician	0.667 (0.400–0.923)	0.600 (0.300–0.900)	<.001
		Physician in training	0.259 (0.091–0.455)	0.400 (0.100–0.750)	<.001
		Nurse	0.313 (0.143–0.480)	0.467 (0.231–0.734)	<.001
	Specificity	Subspecialized physician	0.886 (0.765–0.973)	0.885 (0.769–0.971)	0.914
		Physician in training	0.579 (0.421–0.732)	0.854 (0.730–0.951)	<.001
		Nurse	0.765 (0.667–0.846)	0.893 (0.829–0.955)	<.001
	Sensitivity	Subspecialized physician	0.727 (0.444–1.000)	0.556 (0.250–0.834)	<.001
		Physician in training	0.600 (0.266–0.889)	0.385 (0.100–0.750)	<.001
		Nurse	0.450 (0.231–0.692)	0.389 (0.176–0.611)	<.001
	F1 Score	Subspecialized physician	0.694 (0.421–0.889)	0.571 (0.286–0.800)	<.001
		Physician in training	0.357 (0.143–0.564)	0.385 (0.105–0.667)	<.001
		Nurse	0.367 (0.182–0.528)	0.429 (0.207–0.609)	<.001
	NPV	Subspecialized physician	0.912 (0.800–1.000)	0.861 (0.737–0.969)	<.001
		Physician in training	0.852 (0.714–0.967)	0.853 (0.738–0.952)	0.506
		Nurse	0.855 (0.765–0.931)	0.862 (0.787–0.929)	<.001
Delirium	Accuracy	Subspecialized physician	0.774 (0.644–0.903)	0.774 (0.613–0.903)	0.659
		Physician in training	0.732 (0.585–0.854)	0.805 (0.683–0.927)	<.001
		Nurse	0.740 (0.630–0.836)	0.767 (0.671–0.863)	<.001
	Balanced Accuracy	Subspecialized physician	0.743 (0.577–0.900)	0.742 (0.560–0.905)	0.726
		Physician in training	0.689 (0.521–0.848)	0.795 (0.629–0.926)	<.001
		Nurse	0.727 (0.613–0.831)	0.735 (0.626–0.841)	<.001

	PPV	Subspecialized physician	0.812 (0.632–0.955)	0.810 (0.625–0.955)	0.897
		Physician in training	0.826 (0.655–0.958)	0.889 (0.741–1.000)	<.001
		Nurse	0.826 (0.703–0.925)	0.818 (0.702–0.923)	0.003
	Specificity	Subspecialized physician	0.636 (0.300–0.909)	0.636 (0.300–0.909)	0.787
		Physician in training	0.583 (0.273–0.867)	0.750 (0.454–1.000)	<.001
		Nurse	0.681 (0.500–0.850)	0.640 (0.444–0.826)	<.001
	Sensitivity	Subspecialized physician	0.857 (0.696–1.000)	0.854 (0.684–1.000)	0.408
		Physician in training	0.800 (0.630–0.931)	0.833 (0.679–0.962)	<.001
		Nurse	0.776 (0.646–0.891)	0.837 (0.717–0.930)	<.001
	F1 Score	Subspecialized physician	0.829 (0.686–0.933)	0.829 (0.683–0.939)	0.435
		Physician in training	0.807 (0.678–0.909)	0.857 (0.739–0.941)	<.001
		Nurse	0.800 (0.698–0.879)	0.825 (0.738–0.900)	<.001
	NPV	Subspecialized physician	0.714 (0.400–1.000)	0.700 (0.400–1.000)	0.806
		Physician in training	0.538 (0.273–0.833)	0.645 (0.385–0.889)	<.001
		Nurse	0.607 (0.423–0.800)	0.667 (0.458–0.846)	<.001

eTable 12. Prospective comparison of clinician and AI prediction performance for the regression outcomes. Results are stratified by clinician type: subspecialized physicians, physicians in training, and nurses. Statistical significance was assessed using two-sample Wilcoxon rank-sum tests. CI: confidence interval, LOS: length of stay, MAE: mean absolute error, RMSE: root mean squared error.

	Metric	Group	Clinician, median (95% CI)	AI, median (95% CI)	P Value
ICU LOS	MAE	Subspecialized physician	4.996 (3.650–6.650)	5.213 (4.314–6.557)	<0.001
		Physician in training	5.208 (3.571–7.307)	6.436 (4.997–8.500)	<0.001
		Nurse	4.209 (3.345–5.216)	5.421 (4.732–6.191)	<0.001
	RMSE	Subspecialized physician	7.613 (5.534–9.998)	6.719 (4.993–9.286)	<0.001
		Physician in training	9.295 (5.087–13.839)	9.641 (6.108–14.277)	<0.001
		Nurse	7.512 (5.562–9.418)	7.248 (5.909–8.559)	<0.001
	R ²	Subspecialized physician	−0.232 (−0.807–0.154)	0.025 (−0.467–0.420)	<0.001
		Physician in training	0.022 (−0.520–0.317)	−0.056 (−0.551–0.061)	<0.001
		Nurse	−0.089 (−0.575–0.213)	−0.014 (−0.382–0.190)	<0.001
Hospital LOS	MAE	Subspecialized physician	42.646 (28.483–61.359)	37.798 (25.583–54.494)	<0.001
		Physician in training	37.036 (23.333–52.403)	39.574 (27.727–53.825)	<0.001
		Nurse	33.611 (24.712–44.523)	34.894 (26.809–44.947)	<0.001
	RMSE	Subspecialized physician	76.854 (45.653–109.661)	66.271 (37.628–99.160)	<0.001
		Physician in training	72.464 (45.615–95.056)	70.102 (45.741–92.113)	<0.001
		Nurse	67.643 (50.857–85.955)	64.545 (48.228–83.089)	<0.001
	R ²	Subspecialized physician	−0.334 (−0.572–−0.147)	0.010 (−0.143–0.195)	<0.001
		Physician in training	−0.029 (−0.232–0.223)	0.030 (−0.073–0.122)	<0.001
		Nurse	−0.053 (−0.157–0.092)	0.046 (−0.029–0.132)	<0.001

eTable 13. Inter-rater agreement between clinicians for the classification outcomes in the retrospective setting. The one-clinician group is not shown since this analysis requires at least 2 clinicians predicting the same patient. AKI: acute kidney injury, SD: standard deviation.

Outcome	Group	Cohen's Kappa, mean (SD)
In-hospital mortality	2 clinicians	0.2 (0.13)
	7 clinicians	0.28 (0.24)
30-Day post-discharge mortality	2 clinicians	0.06 (0.09)
	7 clinicians	0.24 (0.25)
AKI	2 clinicians	0.42 (0.17)
	7 clinicians	0.41 (0.19)
Delirium	2 clinicians	0.23 (0.3)
	7 clinicians	0.23 (0.17)

eTable 14. Inter-rater agreement between clinicians for the regression outcomes in the retrospective setting. The one-clinician group is not shown since this analysis requires at least 2 clinicians predicting the same patient. ICC: intraclass correlation coefficient, ICU: intensive care unit, LOS: length of stay, SD: standard deviation.

Outcome	Group	ICC(3,1), mean (SD)
ICU LOS	2 clinicians	0.3 (0.11)
	7 clinicians	0.25 (0.05)
Hospital LOS	2 clinicians	0.03 (0.09)
	7 clinicians	0.09 (0.06)

eTable 15. Inter-rater agreement between clinicians for the classification outcomes in the prospective setting. This analysis was stratified by the number of patients predicting the same patient across clinician types. AKI: acute kidney injury, SD: standard deviation.

Outcome	Number of Clinicians per Patient	Number of Patients	Cohen's Kappa (SD)
In-hospital mortality	2 clinicians	68	0.44 (0.17)
	3 clinicians	15	0.00 (0.02)
	4 clinicians	9	0.00 (0.03)
30-day post-discharge mortality	2 clinicians	68	0.62 (0.15)
	3 clinicians	15	-0.05 (0.04)
	4 clinicians	9	0.00 (0.03)
AKI	2 clinicians	67	0.54 (0.12)
	3 clinicians	15	0.26 (0.14)
	4 clinicians	9	0.49 (0.26)
Delirium	2 clinicians	67	0.29 (0.12)
	3 clinicians	15	0.13 (0.17)
	4 clinicians	9	0.11 (0.21)

eTable 16. Inter-rater agreement between clinicians for the regression outcomes in the prospective setting. This analysis was stratified by the number of patients predicting the same patient across clinician types. ICC: intraclass correlation coefficient, ICU: intensive care unit, LOS: length of stay, SD: standard deviation.

Outcome	Number of Clinicians per Patient	Number of Patients	ICC (1,1), mean (SD)
ICU LOS	2 clinicians	67	0.41 (0.09)
	3 clinicians	15	0.05 (0.12)
	4 clinicians	8	0.10 (0.28)
Hospital LOS	2 clinicians	66	0.63 (0.11)
	3 clinicians	14	0.15 (0.08)
	4 clinicians	8	-0.03 (0.19)

eTable 17. Inter-rater agreement between clinicians and AI for the classification outcomes in the retrospective setting. In the groups with more than one clinician, clinician predictions were averaged and compared against corresponding AI predictions. AKI: acute kidney injury, SD: standard deviation.

Outcome	Group	Cohen's Kappa, mean (SD)
In-hospital Mortality	1 clinician	0.343 (0.049)
	2 clinicians	0.334 (0.066)
	7 clinicians	0.702 (0.152)
30-day post-discharge mortality	1 clinician	0.294 (0.056)
	2 clinicians	0.189 (0.069)
	7 clinicians	0.696 (0.182)
AKI	1 clinician	0.561 (0.044)
	2 clinicians	0.353 (0.060)
	7 clinicians	0.267 (0.137)
Delirium	1 clinician	0.251 (0.054)
	2 clinicians	0.394 (0.065)
	7 clinicians	0.677 (0.150)

eTable 18. Inter-rater agreement between clinicians and AI for the regression outcomes in the retrospective setting. In the groups with more than one clinician, clinician predictions were averaged and compared against corresponding AI predictions. ICC: intraclass correlation coefficient, ICU: intensive care unit, LOS: length of stay, SD: standard deviation.

Outcome	Group	ICC(3,1), mean (SD)
ICU LOS	1 clinician	0.255 (0.031)
	2 clinicians	0.499 (0.041)
	7 clinicians	0.624 (0.079)
Hospital LOS	1 clinician	0.165 (0.055)
	2 clinicians	0.288 (0.047)
	7 clinicians	0.451 (0.109)

eTable 19. Inter-rater agreement between clinicians and AI for the classification outcomes in the prospective setting. This analysis was stratified by the number of patients predicting the same patient across clinician types. AKI: acute kidney injury, SD: standard deviation.

Outcome	Number of Clinicians per Patient	Number of Patients	Cohen's Kappa, mean (SD)
Hospital mortality	1 clinician	181	0.349 (0.082)
	2 clinicians	124	0.363 (0.116)
	3 clinicians	48	-0.041 (0.039)
30-day post-discharge mortality	1 clinician	181	0.425 (0.092)
	2 clinicians	124	0.447 (0.120)
	3 clinicians	48	0.064 (0.084)
AKI	1 clinician	181	0.412 (0.082)
	2 clinicians	123	0.384 (0.101)
	3 clinicians	48	0.287 (0.143)
Delirium	1 clinician	181	0.196 (0.049)
	2 clinicians	123	0.188 (0.077)
	3 clinicians	48	0.033 (0.078)

eTable 20. Inter-rater agreement between clinicians and AI for the regression outcomes in the prospective setting. This analysis was stratified by the number of patients predicting the same patient across clinician types. ICC: intraclass correlation coefficient, ICU: intensive care unit, LOS: length of stay, SD: standard deviation.

Outcome	Group	Number of Patients	ICC (1,1), mean (SD)
ICU LOS	1 clinician	149	0.243 (0.049)
	2 clinicians	52	0.184 (0.078)
	3 clinicians	12	0.104 (0.090)
Hospital LOS	1 clinician	149	0.193 (0.073)
	2 clinicians	52	0.157 (0.062)
	3 clinicians	12	-0.032 (0.055)

eTable 21. Likert scale threshold sensitivity analysis results for the retrospective comparison of clinician and AI prediction performances for the classification outcomes. Results are stratified by the number of clinicians providing predictions: one, two, and seven clinicians. Statistical significance was assessed using two-sample Wilcoxon rank-sum tests. AKI: acute kidney injury, CI: confidence interval, NPV: negative predictive value, PPV: positive predictive value.

Outcome	Metric	Group	Clinicians, median (95% CI)	AI, median (95% CI)	P Value
In-hospital Mortality	Accuracy	One clinician	0.630 (0.589–0.668)	0.864 (0.836–0.891)	<.001
		Two clinicians	0.616 (0.570–0.669)	0.836 (0.798–0.871)	<.001
		Seven clinicians	0.520 (0.360–0.660)	0.760 (0.640–0.880)	<.001
	Balanced Accuracy	One clinician	0.660 (0.607–0.709)	0.732 (0.675–0.785)	<.001
		Two clinicians	0.633 (0.565–0.702)	0.704 (0.641–0.777)	<.001
		Seven clinicians	0.669 (0.491–0.786)	0.538 (0.409–0.744)	<.001
	PPV	One clinician	0.252 (0.200–0.307)	0.575 (0.459–0.673)	<.001
		Two clinicians	0.236 (0.173–0.305)	0.475 (0.355–0.593)	<.001
		Seven clinicians	0.222 (0.077–0.400)	0.231 (0.000–0.571)	0.204
	Specificity	One clinician	0.614 (0.573–0.656)	0.925 (0.899–0.947)	<.001
		Two clinicians	0.607 (0.558–0.663)	0.896 (0.861–0.926)	<.001
		Seven clinicians	0.450 (0.308–0.610)	0.860 (0.756–0.952)	<.001
	Sensitivity	One clinician	0.706 (0.608–0.800)	0.538 (0.432–0.647)	<.001
		Two clinicians	0.661 (0.525–0.784)	0.514 (0.386–0.655)	<.001
		Seven clinicians	0.889 (0.571–1.000)	0.222 (0.000–0.600)	<.001
	F1 Score	One clinician	0.372 (0.306–0.434)	0.555 (0.460–0.639)	<.001
		Two clinicians	0.347 (0.266–0.428)	0.493 (0.379–0.602)	<.001
		Seven clinicians	0.357 (0.143–0.558)	0.222 (0.000–0.526)	<.001
	NPV	One clinician	0.919 (0.888–0.948)	0.915 (0.889–0.939)	<.001

		Two clinicians	0.907 (0.866–0.944)	0.910 (0.876–0.940)	0.001
		Seven clinicians	0.952 (0.833–1.000)	0.857 (0.750–0.954)	<.001
30-Day Post-Discharge Mortality	Accuracy	One clinician	0.608 (0.569–0.650)	0.869 (0.839–0.896)	<.001
		Two clinicians	0.674 (0.626–0.719)	0.861 (0.823–0.893)	<.001
		Seven clinicians	0.520 (0.380–0.660)	0.800 (0.680–0.900)	<.001
	Balanced Accuracy	One clinician	0.654 (0.605–0.706)	0.694 (0.639–0.748)	<.001
		Two clinicians	0.640 (0.577–0.706)	0.690 (0.627–0.753)	<.001
		Seven clinicians	0.629 (0.459–0.759)	0.565 (0.464–0.718)	<.001
	PPV	One clinician	0.244 (0.196–0.299)	0.618 (0.500–0.745)	<.001
		Two clinicians	0.269 (0.198–0.341)	0.590 (0.439–0.737)	<.001
		Seven clinicians	0.258 (0.107–0.432)	0.500 (0.000–1.000)	<.001
	Specificity	One clinician	0.588 (0.545–0.633)	0.949 (0.929–0.970)	<.001
		Two clinicians	0.690 (0.636–0.738)	0.941 (0.912–0.965)	<.001
		Seven clinicians	0.446 (0.300–0.610)	0.951 (0.878–1.000)	<.001
	Sensitivity	One clinician	0.719 (0.634–0.812)	0.440 (0.330–0.547)	<.001
		Two clinicians	0.590 (0.471–0.712)	0.439 (0.315–0.564)	<.001
		Seven clinicians	0.818 (0.500–1.000)	0.182 (0.000–0.500)	<.001
	F1 Score	One clinician	0.364 (0.302–0.427)	0.513 (0.407–0.604)	<.001
		Two clinicians	0.371 (0.284–0.453)	0.500 (0.379–0.615)	<.001
		Seven clinicians	0.389 (0.182–0.583)	0.267 (0.000–0.571)	<.001
	NPV	One clinician	0.919 (0.887–0.949)	0.900 (0.875–0.925)	<.001
		Two clinicians	0.897 (0.856–0.932)	0.897 (0.862–0.927)	0.813
		Seven clinicians	0.909 (0.762–1.000)	0.830 (0.711–0.918)	<.001

AKI	Accuracy	One clinician	0.533 (0.484–0.587)	0.768 (0.719–0.809)	<.001
		Two clinicians	0.423 (0.365–0.486)	0.747 (0.693–0.801)	<.001
		Seven clinicians	0.485 (0.303–0.667)	0.758 (0.606–0.879)	<.001
	Balanced Accuracy	One clinician	0.611 (0.567–0.658)	0.708 (0.653–0.761)	<.001
		Two clinicians	0.509 (0.446–0.574)	0.667 (0.602–0.735)	<.001
		Seven clinicians	0.673 (0.586–0.762)	0.523 (0.417–0.682)	<.001
	PPV	One clinician	0.379 (0.316–0.443)	0.637 (0.535–0.731)	<.001
		Two clinicians	0.297 (0.228–0.365)	0.585 (0.440–0.705)	<.001
		Seven clinicians	0.291 (0.111–0.478)	0.333 (0.000–1.000)	0.576
	Specificity	One clinician	0.411 (0.354–0.471)	0.860 (0.816–0.903)	<.001
		Two clinicians	0.306 (0.244–0.376)	0.863 (0.807–0.911)	<.001
		Seven clinicians	0.346 (0.172–0.524)	0.926 (0.808–1.000)	<.001
	Sensitivity	One clinician	0.811 (0.732–0.891)	0.551 (0.464–0.650)	<.001
		Two clinicians	0.711 (0.603–0.818)	0.473 (0.358–0.592)	<.001
		Seven clinicians	1.000 (1.000–1.000)	0.125 (0.000–0.429)	<.001
	F1 Score	One clinician	0.516 (0.451–0.579)	0.591 (0.510–0.670)	<.001
		Two clinicians	0.418 (0.338–0.490)	0.520 (0.404–0.624)	<.001
		Seven clinicians	0.451 (0.200–0.647)	0.182 (0.000–0.500)	<.001
	NPV	One clinician	0.832 (0.764–0.901)	0.815 (0.767–0.860)	<.001
		Two clinicians	0.727 (0.620–0.825)	0.798 (0.740–0.856)	<.001
		Seven clinicians	1.000 (1.000–1.000)	0.800 (0.636–0.931)	<.001
Delirium	Accuracy	One clinician	0.648 (0.594–0.698)	0.800 (0.752–0.838)	<.001
		Two clinicians	0.706 (0.642–0.765)	0.784 (0.730–0.838)	<.001

		Seven clinicians	0.520 (0.320–0.720)	0.760 (0.599–0.920)	<.001
	Balanced Accuracy	One clinician	0.626 (0.582–0.669)	0.798 (0.751–0.837)	<.001
		Two clinicians	0.652 (0.594–0.711)	0.778 (0.721–0.837)	<.001
		Seven clinicians	0.536 (0.500–0.615)	0.763 (0.566–0.920)	<.001
	PPV	One clinician	0.622 (0.558–0.680)	0.810 (0.750–0.861)	<.001
		Two clinicians	0.690 (0.617–0.762)	0.830 (0.761–0.896)	<.001
		Seven clinicians	0.500 (0.318–0.696)	0.750 (0.462–1.000)	<.001
	Specificity	One clinician	0.351 (0.273–0.428)	0.775 (0.704–0.836)	<.001
		Two clinicians	0.377 (0.274–0.481)	0.753 (0.654–0.845)	<.001
		Seven clinicians	0.071 (0.000–0.231)	0.769 (0.500–1.000)	<.001
	Sensitivity	One clinician	0.900 (0.851–0.944)	0.823 (0.764–0.876)	<.001
		Two clinicians	0.927 (0.874–0.969)	0.805 (0.733–0.868)	<.001
		Seven clinicians	1.000 (1.000–1.000)	0.769 (0.500–1.000)	<.001
	F1 Score	One clinician	0.735 (0.685–0.780)	0.815 (0.771–0.857)	<.001
		Two clinicians	0.791 (0.735–0.840)	0.817 (0.762–0.865)	<.001
		Seven clinicians	0.667 (0.483–0.821)	0.750 (0.500–0.917)	<.001
	NPV	One clinician	0.750 (0.644–0.851)	0.789 (0.724–0.850)	<.001
		Two clinicians	0.775 (0.628–0.898)	0.720 (0.624–0.812)	<.001
		Seven clinicians	1.000 (0.000–1.000)	0.786 (0.545–1.000)	<.001

eTable 22. Likert scale threshold sensitivity analysis results for the prospective comparison of clinician and AI prediction performances for the classification outcomes. Results are stratified by clinician type: subspecialized physicians, physicians in training, and nurses. Statistical significance was assessed using two-sample Wilcoxon rank-sum tests. AKI: acute kidney injury, CI: confidence interval, NPV: negative predictive value. AKI: acute kidney injury, CI: confidence interval, NPV: negative predictive value, PPV: positive predictive value.

Outcome	Metric	Group	Clinician, median (95% CI)	AI, median (95% CI)	P Value
In-hospital Mortality	Accuracy	Subspecialized	0.763 (0.632–0.895)	0.789 (0.658–0.921)	<0.001
		Trainee	0.717 (0.600–0.833)	0.917 (0.833–0.983)	<0.001
		Nurse	0.718 (0.636–0.809)	0.791 (0.718–0.864)	<0.001
	Balanced Accuracy	Subspecialized	0.670 (0.527–0.832)	0.761 (0.611–0.904)	<0.001
		Trainee	0.600 (0.502–0.705)	0.917 (0.825–0.978)	<0.001
		Nurse	0.632 (0.553–0.721)	0.782 (0.701–0.858)	<0.001
	PPV	Subspecialized	0.781 (0.618–0.912)	0.852 (0.696–0.964)	<0.001
		Trainee	0.722 (0.593–0.830)	0.951 (0.868–1.000)	<0.001
		Nurse	0.719 (0.626–0.816)	0.857 (0.757–0.928)	<0.001
	Specificity	Subspecialized	0.412 (0.143–0.700)	0.667 (0.384–0.923)	<0.001
		Trainee	0.250 (0.067–0.445)	0.905 (0.750–1.000)	<0.001
		Nurse	0.333 (0.195–0.500)	0.744 (0.600–0.868)	<0.001
	Sensitivity	Subspecialized	0.926 (0.808–1.000)	0.852 (0.692–0.966)	<0.001
		Trainee	0.953 (0.879–1.000)	0.930 (0.844–1.000)	<0.001
		Nurse	0.931 (0.859–0.985)	0.818 (0.729–0.900)	<0.001
	F1 Score	Subspecialized	0.844 (0.723–0.935)	0.846 (0.727–0.947)	0.139
		Trainee	0.821 (0.723–0.896)	0.940 (0.873–0.987)	<0.001
		Nurse	0.812 (0.737–0.874)	0.836 (0.766–0.893)	<0.001
	NPV	Subspecialized	0.714 (0.333–1.000)	0.667 (0.384–0.917)	<0.001
		Trainee	0.727 (0.333–1.000)	0.870 (0.695–1.000)	<0.001
		Nurse	0.727 (0.500–0.917)	0.694 (0.553–0.816)	<0.001

30-day post-discharge mortality	Accuracy	Subspecialized	0.649 (0.526–0.755)	0.772 (0.667–0.877)	<0.001
		Trainee	0.384 (0.274–0.493)	0.795 (0.699–0.890)	<0.001
		Nurse	0.519 (0.444–0.593)	0.796 (0.728–0.858)	<0.001
	Balanced Accuracy	Subspecialized	0.697 (0.577–0.810)	0.704 (0.576–0.828)	0.125
		Trainee	0.569 (0.443–0.669)	0.721 (0.577–0.868)	<0.001
		Nurse	0.641 (0.558–0.715)	0.683 (0.589–0.781)	<0.001
	PPV	Subspecialized	0.448 (0.281–0.625)	0.643 (0.385–0.900)	<0.001
		Trainee	0.204 (0.096–0.314)	0.444 (0.211–0.706)	<0.001
		Nurse	0.253 (0.173–0.337)	0.452 (0.289–0.640)	<0.001
	Specificity	Subspecialized	0.579 (0.421–0.721)	0.878 (0.767–0.974)	<0.001
		Trainee	0.281 (0.167–0.400)	0.836 (0.733–0.926)	<0.001
		Nurse	0.447 (0.373–0.531)	0.864 (0.800–0.920)	<0.001
	Sensitivity	Subspecialized	0.824 (0.615–1.000)	0.529 (0.300–0.750)	<0.001
		Trainee	0.850 (0.636–1.000)	0.609 (0.333–0.889)	<0.001
		Nurse	0.836 (0.687–0.963)	0.500 (0.333–0.690)	<0.001
	F1 Score	Subspecialized	0.577 (0.409–0.732)	0.581 (0.348–0.769)	0.781
		Trainee	0.328 (0.172–0.466)	0.500 (0.276–0.720)	<0.001
		Nurse	0.388 (0.278–0.487)	0.475 (0.314–0.630)	<0.001
	NPV	Subspecialized	0.889 (0.741–1.000)	0.814 (0.698–0.917)	<0.001
		Trainee	0.900 (0.733–1.000)	0.911 (0.823–0.981)	0.003
		Nurse	0.924 (0.852–0.984)	0.885 (0.829–0.937)	<0.001
AKI	Accuracy	Subspecialized	0.738 (0.643–0.833)	0.810 (0.726–0.893)	<0.001
		Trainee	0.779 (0.702–0.856)	0.750 (0.663–0.837)	<0.001
		Nurse	0.727 (0.671–0.788)	0.788 (0.732–0.840)	<0.001
	Balanced Accuracy	Subspecialized	0.759 (0.660–0.850)	0.754 (0.646–0.848)	0.003
		Trainee	0.770 (0.676–0.859)	0.673 (0.576–0.769)	<0.001
		Nurse	0.698 (0.628–0.764)	0.667 (0.600–0.736)	<0.001

	PPV	Subspecialized	0.567 (0.412–0.719)	0.765 (0.571–0.917)	<0.001
		Trainee	0.600 (0.447–0.750)	0.600 (0.393–0.789)	0.92
		Nurse	0.489 (0.379–0.600)	0.658 (0.500–0.806)	<0.001
	Specificity	Subspecialized	0.704 (0.586–0.818)	0.915 (0.833–0.980)	<0.001
		Trainee	0.792 (0.696–0.879)	0.864 (0.773–0.937)	<0.001
		Nurse	0.766 (0.697–0.825)	0.924 (0.880–0.961)	<0.001
	Sensitivity	Subspecialized	0.818 (0.667–0.955)	0.593 (0.400–0.769)	<0.001
		Trainee	0.750 (0.586–0.889)	0.485 (0.310–0.667)	<0.001
		Nurse	0.630 (0.500–0.745)	0.413 (0.286–0.538)	<0.001
	F1 Score	Subspecialized	0.667 (0.523–0.783)	0.667 (0.489–0.800)	0.57
		Trainee	0.667 (0.531–0.779)	0.533 (0.367–0.677)	<0.001
		Nurse	0.550 (0.446–0.641)	0.505 (0.371–0.623)	<0.001
	NPV	Subspecialized	0.891 (0.792–0.976)	0.831 (0.723–0.912)	<0.001
		Trainee	0.882 (0.800–0.956)	0.800 (0.707–0.883)	<0.001
		Nurse	0.852 (0.796–0.902)	0.814 (0.758–0.867)	<0.001
Delirium	Accuracy	Subspecialized	0.679 (0.571–0.774)	0.774 (0.679–0.857)	<0.001
		Trainee	0.692 (0.606–0.788)	0.731 (0.644–0.817)	<0.001
		Nurse	0.710 (0.654–0.762)	0.779 (0.727–0.831)	<0.001
	Balanced Accuracy	Subspecialized	0.701 (0.583–0.801)	0.722 (0.615–0.829)	<0.001
		Trainee	0.697 (0.604–0.793)	0.668 (0.571–0.773)	<0.001
		Nurse	0.706 (0.636–0.770)	0.651 (0.577–0.723)	<0.001
	PPV	Subspecialized	0.447 (0.281–0.600)	0.579 (0.389–0.773)	<0.001
		Trainee	0.488 (0.348–0.631)	0.548 (0.368–0.727)	<0.001
		Nurse	0.419 (0.321–0.523)	0.522 (0.370–0.675)	<0.001
	Specificity	Subspecialized	0.657 (0.536–0.767)	0.836 (0.746–0.922)	<0.001
		Trainee	0.684 (0.580–0.791)	0.824 (0.729–0.900)	<0.001
		Nurse	0.714 (0.645–0.778)	0.888 (0.838–0.932)	<0.001

	Sensitivity	Subspecialized	0.739 (0.555–0.909)	0.609 (0.400–0.818)	<0.001
		Trainee	0.714 (0.555–0.870)	0.517 (0.333–0.696)	<0.001
		Nurse	0.698 (0.569–0.815)	0.418 (0.283–0.554)	<0.001
	F1 Score	Subspecialized	0.556 (0.392–0.692)	0.592 (0.419–0.739)	<0.001
		Trainee	0.575 (0.441–0.696)	0.531 (0.356–0.667)	<0.001
		Nurse	0.521 (0.422–0.620)	0.462 (0.329–0.589)	<0.001
	NPV	Subspecialized	0.875 (0.761–0.957)	0.852 (0.746–0.935)	<0.001
		Trainee	0.852 (0.759–0.939)	0.802 (0.701–0.889)	<0.001
		Nurse	0.888 (0.836–0.935)	0.837 (0.786–0.886)	<0.001