Supplementary File

Performance of Machine Learning Models for Predicting Post-Induction Hypotension: A Systematic Review and Meta-Analysis

Supplementary Table S1. Search strategies for various databases.

Database	Search strategy and details	Hits
PubMed	(("machine learning"[Title/Abstract] OR "deep learning"[Title/Abstract] OR "artificial intelligence"[Title/Abstract] OR "neural network"[Title/Abstract] OR "random forest"[Title/Abstract] OR "support vector machine"[Title/Abstract] OR "XGBoost"[Title/Abstract] OR "AI"[Title/Abstract] OR "supervised learning"[Title/Abstract] OR "gradient boosting"[Title/Abstract] OR "decision tree"[Title/Abstract] OR "predictive analytics"[Title/Abstract] OR "computer assisted decision making"[Title/Abstract] OR "algorithm"[Title/Abstract]) AND ("hypotension"[Title/Abstract] OR "hypotension"[MeSH Terms] OR "low blood pressure"[Title/Abstract] OR "arterial pressure"[Title/Abstract] OR "mean arterial pressure"[Title/Abstract] OR "MAP"[Title/Abstract] OR "hemodynamic instability"[Title/Abstract] OR "hemodynamic depression"[Title/Abstract] OR "blood pressure"[MeSH Terms] OR "vasopressor"[Title/Abstract] OR "hypotensive episode"[Title/Abstract]) AND ("anesthesia induction"[Title/Abstract] OR "induction of anesthesia"[Title/Abstract] OR "anaesthesia induction"[Title/Abstract] OR "induction of anaesthesia"[Title/Abstract] OR "post-induction"[Title/Abstract] OR "postinduction"[Title/Abstract] OR "peri-induction"[Title/Abstract] OR "periinduction"[Title/Abstract] OR "after induction"[Title/Abstract] OR "post-intubation"[Title/Abstract] OR "postintubation"[Title/Abstract] OR "induct*"[Title/Abstract]) AND ("anesthesia"[Title/Abstract] OR "anaesthesia"[Title/Abstract] OR "anesthesiology"[Title/Abstract] OR "anaesthesiology"[Title/Abstract] OR "general anesthesia"[Title/Abstract] OR "general anaesthesia"[Title/Abstract] OR "surgery"[Title/Abstract] OR "surgical procedures, operative"[MeSH Terms] OR "intraoperative"[Title/Abstract] OR "perioperative"[Title/Abstract] OR "operating room"[Title/Abstract] OR "preoperative"[Title/Abstract])) NOT "review"[Publication Type]	80
Embase	('machine learning':ti,ab,kw OR 'deep learning':ti,ab,kw OR 'artificial intelligence':ti,ab,kw OR 'neural network':ti,ab,kw OR 'random forest':ti,ab,kw OR 'support vector machine':ti,ab,kw OR 'xgboost':ti,ab,kw OR 'ai':ti,ab,kw OR 'supervised learning':ti,ab,kw OR 'gradient boosting':ti,ab,kw OR 'decision tree':ti,ab,kw OR 'predictive analytics':ti,ab,kw OR 'computer assisted decision making':ti,ab,kw OR 'algorithm':ti,ab,kw) AND ('hypotension':ti,ab,kw OR 'low blood pressure':ti,ab,kw OR 'arterial pressure':ti,ab,kw OR 'mean arterial pressure':ti,ab,kw OR 'map':ti,ab,kw OR 'hemodynamic instability':ti,ab,kw OR 'hemodynamic depression':ti,ab,kw OR 'blood pressure':ti,ab,kw OR 'vasopressor':ti,ab,kw OR 'hypotensive episode':ti,ab,kw) AND ('anesthesia induction':ti,ab,kw OR 'induction of anesthesia':ti,ab,kw OR 'anaesthesia induction':ti,ab,kw OR 'induction of anaesthesia':ti,ab,kw OR 'post-induction':ti,ab,kw OR 'postinduction':ti,ab,kw OR 'peri-induction':ti,ab,kw OR 'periinduction':ti,ab,kw OR 'after induction':ti,ab,kw OR 'post-intubation':ti,ab,kw OR 'postintubation':ti,ab,kw OR 'tracheal intubation':ti,ab,kw OR 'induct*':ti,ab,kw) AND ('anesthesia':ti,ab,kw OR 'anaesthesia':ti,ab,kw OR 'anesthesiology':ti,ab,kw OR 'anaesthesiology':ti,ab,kw OR 'general anesthesia':ti,ab,kw OR 'general anaesthesia':ti,ab,kw OR 'surgery':ti,ab,kw OR 'surgical procedures, operative':ti,ab,kw OR 'intraoperative':ti,ab,kw OR 'perioperative':ti,ab,kw OR 'operating room':ti,ab,kw OR 'preoperative':ti,ab,kw) Reviews, conferences, and letters were excluded by filter.	166
Scopus	(TITLE-ABS-KEY ("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network" OR "random forest" OR "support vector machine" OR "XGBoost" OR "AI" OR "supervised learning" OR "gradient boosting" OR "decision tree" OR "predictive analytics" OR "computer assisted decision making" OR "algorithm")) AND (TITLE-ABS-KEY ("hypotension" OR "low blood pressure" OR "arterial pressure" OR "mean arterial pressure" OR "MAP" OR "hemodynamic instability" OR "hemodynamic depression" OR "vasopressor" OR "hypotensive episode" OR "blood pressure")) AND (TITLE-ABS-KEY ("anesthesia induction" OR "induction of anesthesia" OR "anaesthesia induction" OR "induction of anaesthesia" OR "post-induction" OR "postinduction" OR "peri-induction" OR "periinduction" OR "after induction" OR "post-intubation" OR "postintubation" OR "induction" OR "tracheal intubation")) AND (TITLE-ABS-KEY ("anesthesia" OR "anaesthesia" OR "anesthesiology" OR "anaesthesiology" OR "general anesthesia" OR "general anaesthesia" OR "surgery" OR "intraoperative" OR "perioperative" OR "operating room" OR "preoperative" OR "surgical procedures")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English"))	223
Web of Science	("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network" OR "random forest" OR "support vector machine" OR "XGBoost" OR "AI" OR "supervised learning" OR "gradient boosting" OR "decision tree" OR "predictive analytics" OR "computer assisted decision making" OR "algorithm") AND ("hypotension" OR "low blood pressure" OR "arterial pressure" OR "mean arterial pressure" OR "MAP" OR "hemodynamic instability" OR "hemodynamic depression" OR "vasopressor" OR "hypotensive episode" OR "blood pressure") AND ("anesthesia induction" OR "induction of anesthesia" OR "anaesthesia induction" OR "post-induction" OR "postinduction" OR "peri-induction" OR "periinduction" OR "after induction" OR "post-intubation" OR "postintubation" OR "induction" OR "tracheal intubation") AND ("anesthesia" OR "anaesthesia" OR "anesthesiology" OR "anaesthesiology" OR "general anesthesia" OR "general anaesthesia" OR "surgery" OR "intraoperative" OR "perioperative" OR "operating room" OR "preoperative" OR "surgical procedures") Article type and English language were chosen from the filters.	127
Total records retrieved from all databases		596

Duplicate records removed	216
Records remaining after duplicate removal	**380**
Date: 4 October, 2025	

Supplementary Table S2. PICO framework

Population (P)	Adult patients undergoing surgical procedures under general anesthesia.
Intervention (I)	Application of machine learning–based models for the prediction of post-induction hypotension during general anesthesia.
Comparison (C)	No comparator required
Outcome (O)	Predictive performance of machine learning models, including discrimination, calibration, accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUROC), and other reported performance metrics.

Supplementary Table S3. Extracted variables (based on the CHARMS checklist)

Domain	Items extracted
Study	First author; year; country; design; data source and period; number of centers; reporting guideline cited (TRIPOD/TRIPOD+AI)
Participants	Inclusion and exclusion criteria; stated age range; age; sex; ASA physical status; comorbidities; type of surgery and anesthesia; total sample size; number with and without PIH
Outcome	PIH definition (threshold; absolute or relative; single reading or sustained); outcome window; blood pressure measurement method and frequency
Predictors	Candidate and final predictors; number of features; feature-selection method; timing of predictor measurement relative to the outcome window (tier 1: pre-induction only; tier 2: plus induction drug doses; tier 3: post-induction hemodynamic data)
Model development	Algorithms compared; selected (best/final) model and selection rule; handling of missing data; class-imbalance handling (e.g., SMOTE, down-sampling); hyperparameter tuning
Validation	Apparent; (repeated) k-fold cross-validation; bootstrap; random split (held-out test set); temporal validation; geographically external validation; size and events of each evaluation set
Performance	AUC with 95% CI and the evaluation set it refers to; sensitivity; specificity; PPV; NPV; accuracy; F1; classification threshold and threshold-selection method; calibration (plot, slope/intercept, Hosmer-Lemeshow, Brier score); decision-curve analysis; clinical-impact assessment
Interpretability	Variable importance, SHAP or other explanation methods

Supplementary Table S4. Risk of bias and applicability (PROBAST)

Study	Participants	Predictors	Outcome	Analysis	Overall risk of bias	Applicability concern	Main reasons for judgments
Lin 2011	Low	Unclear	High	Unclear	High	Low	D2: induction-drug doses recorded during the post-induction (outcome) period. D3: baseline SBP is both a predictor and part of the outcome definition (>30% fall). D4: missing-data handling not reported; calibration by Hosmer-Lemeshow test and plot only; temporal (same-centre) validation with 106 events.
Kang 2020	Low	Low	Low	High	High	Unclear	D4: 126 events for 89 candidate features (EPV ~1.4); data-driven feature selection; repeated cross-validation only (no independent test set); calibration not assessed. Applicability: single procedure (laparoscopic cholecystectomy) under TIVA.
Lee 2020	Unclear	Low	Low	High	High	Unclear	D1: exclusion of cases with <4 min between induction and intubation. D4: 151 events for 97 candidate features; no AUC; calibration not assessed; cross-validation only. Applicability: single procedure; prediction 1 min before intubation.
Li 2021	Low	High	Unclear	High	High	Low	D2: intra-operative medications up to 10 min after induction (inside the outcome window) used as predictors. D3: outcome window not precisely specified. D4: complete-case exclusion of incomplete records; calibration not assessed.
Yoshimura 2022	Low	Low	Low	High	High	Unclear	D4: 95 candidate features with 670 events (EPV ~7); missing values replaced with 'normal' values; calibration not assessed; reported PPV/NPV inconsistent with sensitivity, specificity and prevalence. Applicability: restricted to patients with recent echocardiography.
Park 2025	Unclear	High	Low	High	High	Unclear	D1: only cases with continuous arterial pressure recording (2,664 of 6,388 excluded). D2: minimum/mean blood pressure between induction and intubation (inside the outcome window) used as predictors. D4: SMOTE without recalibration; calibration not assessed; reported CIs implausibly narrow.
Chen 2025	Low	Low	Low	Unclear	Unclear	Unclear	D4: median imputation of missing data; otherwise held-out test set, calibration curve and Brier score, decision-curve analysis. Applicability: age restriction not reported (open registry includes a small number of minors); data period misreported.
Bian 2025	Low	Low	High	Low	High	Low	D3: ward baseline MAP is both a predictor and part of the outcome definition (relative fall >30%). D4: LASSO selection, multiple imputation, calibration (curve, H-L, Brier), decision-curve analysis and temporal validation.
Katsin 2025	Unclear	Low	Unclear	High	High	Low	D1: 2,456 cases excluded for missing baseline or inadequate BP data. D3: outcome ascertained partly by invasive monitoring started after induction (23.9%). D4: AutoML pipeline; evaluation set ambiguous (CV vs 20% hold-out); calibration not assessed; reported accuracy inconsistent with recall/precision.
Sarullo 2025	Unclear	High	Low	High	High	Low	D1: complete-case cohort; multiple cases per patient treated as independent. D2: propofol/fentanyl doses given within the first 10 min (the outcome window) used as predictors. D4: cross-validation only; clustering ignored; calibration not assessed.

Supplementary Table S5. Detailed characteristics of the included studies

Study	Sample (development / evaluation)	Age, y	Male, %	ASA	Comorbidities	Predictors	Algorithms; validation	Notes
Lin 2011	Training 1,017 / temporal validation 294	46.4 ± 17.4 (validation)	36.4	I–II 9.5%; III–IV 90.5% (as printed in source)	HTN 19.7%; DM 4.1%	18 inputs (11 patient, 2 surgical, 5 anesthetic): age, sex, weight, height, hematocrit, ASA, baseline SBP/DBP/HR, HTN, DM, emergency, position, airway device, thiopental/propofol/etomidate/fentanyl dose	ANN (MLP, 18 hidden nodes); stepwise LR comparator	Age eligibility not stated
Kang 2020	222	53 ± 14	46.9	I 36.5%; II 52.7%; III 10.8%	HTN 32%; DM 20.3%; asthma 7.2%	89 candidate features (EHR + Vital Recorder data from induction to intubation); 20 selected by RFE (feature set C), incl. lowest/mean SBP and MBP before intubation, age, early-PIH frequency/duration	Naive Bayes, LR, RF, ANN; repeated 4-fold CV x1000	Same hospital/database as Lee 2020
Lee 2020	282	54.7 ± 14.1	47.5	I 33.7%; II 54.6%; III 11.7%	HTN 34.4%; DM 22%; asthma 6%	97 candidate features; vital records 4 to 1 min before intubation + EHR; feature set 2 (20 features by RF importance)	RF, XGBoost, CNN, DNN; repeated 10-fold CV x100	No AUC reported

Li 2021	3,030 (training 2,121; test 909)	62 ± 8.9 (whole cohort)	67.2 (whole cohort)	Reported by training/test set	DM, dyslipidemia, recent MI reported	Age, sex, BMI, comorbidities, EuroSCORE I, ASA, hemoglobin, creatinine, bilirubin, preoperative medications, intra-operative medications to 10 min after induction	RF; 10-fold CV within training set; 70/30 split	Cardiac surgery only
Yoshimura 2022	1,956 (training 1,369; test 587)	NR in main text	NR in main text	NR in main text	NR in main text	95 echocardiographic variables reduced to 49 by SHAP	DNN, XGB, LDA, LR; 70/30 split repeated 10 times	Reported PPV/NPV inconsistent with sens/spec and prevalence
Park 2025	3,669 (training 2,568; test 1,101)	58.6 ± 14.4 (test set)	57.8 (test set)	1: 21.1%; 2: 62.8%; 3: 15.3%; 4: 0.6%	HTN, DM reported	46 features: demographics, comorbidities, ASA, surgical details, ECG, PFT, laboratory values, baseline and induction-to-intubation min/mean/max BP and HR, anesthetic agent/propofol dose	RF, LightGBM, AdaBoost, XGB, GBR, ARD; SMOTE; 70/30 split	VitalDB; arterial line cases only
Chen 2025	5,406 (training 3,784; test 1,622)	PIH 58.4 ± 15.5; non-PIH 57.1 ± 14.4	PIH 38.4; non-PIH 51.9	Reported	HTN, DM reported	36 features; 10 most important: baseline DBP, age, sex, surgery type, baseline SBP, platelets, rocuronium use, BUN, albumin, fentanyl use	LR, RF, XGB, NN; 70/30 split + 5-fold CV	VitalDB; age restriction not reported
Bian 2025	938 (development 657; temporal validation 281)	Median 71	52.2	I 2.5%; II 74.7%; III 22.8%	HTN 61.2%; DM 21.0%	Cardiac dysfunction, ward baseline MAP, etomidate use, pre-induction MAP (LASSO + multivariable LR)	LR nomogram; bootstrap (1,000) + temporal validation	Prospective
Katsin 2025	20,309 (training 16,247; validation 4,062)	Median 48.4 (IQR 34.3–68.0)	53.4	Reported	HTN 18.8%; DM 3.9%; CAD 4.0%; AF 3.0%; CHF 1.2%; CKD 1.6%	50 features collected before the post-induction period; 14 contributing >2% (propofol dose, pre-induction SAP/MAP, beta-blocker, etc.)	AutoML (TPOT) -> RF; LR baseline; 5-fold CV	Age reported as median (IQR)
Sarullo 2025	201,000 records (multiple cases per patient)	NR	NR	NR	CCI-based comorbidities	Demographics, comorbidities, procedure-text clusters, preoperative medications, propofol and fentanyl doses in first 10 min	LR, XGB, kNN, NB, RF, LDA, NN, monotonic NN; 5-fold CV	Outcome MAP <65 mmHg (sensitivity analyses at 60 and 55)

Supplementary Table S6. AUC input data and results of the primary and sensitivity analyses

Study	Validation	Family	Evaluation n	Events	Events status	AUC (reported CI)	Hanley-McNeil 95% CI
Lin 2011	Temporal validation	Neural network	294	106	derived (36.1% x 294; Table 1)	0.893 (NR)	0.842–0.929
Kang 2020	Repeated cross-validation	Tree-based	222	126	reported	0.842 (0.736–0.948)	0.785–0.887
Li 2021	Held-out random split (test set)	Tree-based	909	490	derived (53.9% x 909)	0.843 (0.808–0.877)	0.816–0.867
Yoshimura 2022	Held-out random split (test set)	Neural network	587	200	approximate (34% x 587; test-set events not reported)	0.720 (0.670–0.760)	0.672–0.763
Park 2025	Held-out random split (test set)	Tree-based	1,101	672	reported (Table 1 test set)	0.772 (0.771–0.774)	0.744–0.798
Chen 2025	Held-out random split (test set)	Regression-based	1,622	276	approximate (17.0% x 1622; test-set events not reported)	0.740 (0.710–0.770)	0.703–0.774
Bian 2025	Temporal validation	Regression-based	281	142	reported	0.697 (0.635–0.759)	0.633–0.754
Katsin 2025	Cross-validation / held-out (ambiguous)	Tree-based	4,062	991	approximate (24.4% x 4062 validation cohort)	0.732 (NR)	0.712–0.751
Sarullo 2025	5-fold cross-validation	Tree-based	201,000	100,500	assumed balanced (source: 'relatively even'; exact count not in main text)	0.697 (NR)	0.695–0.699

Analysis	k	Pooled AUC	HKSJ 95% CI	Wald 95% CI	95% PI	τ^2 (logit)	I^2, %	Q (df)
Primary: independent cohorts, Hanley-McNeil SE (k=8)	8	0.776	0.700–0.837	0.715–0.827	0.52–0.92	0.196	97.7	142.0 (7)
VitalDB represented by Park 2025 instead of Chen 2025	8	0.780	0.706–0.839	0.721–0.829	0.53–0.92	0.190	97.8	160.6 (7)
All 9 reports (ignores Park/Chen data overlap)	9	0.775	0.711–0.828	0.723–0.820	0.56–0.91	0.165	97.4	165.3 (8)
Temporal/held-out test-set estimates only	5	0.787	0.654–0.879	0.699–0.855	0.38–0.96	0.260	94.6	53.8 (4)

Excluding Sarullo 2025 (clustered cases, CV only)	7	0.787	0.703–0.852	0.723–0.839	0.52–0.93	0.200	94.7	72.2 (6)
Excluding Lin 2011 (largest AUC)	7	0.756	0.689–0.813	0.704–0.802	0.55–0.89	0.116	96.7	110.5 (6)
SE derived from reported 95% CI where available	8	0.771	0.694–0.833	0.709–0.823	0.52–0.91	0.186	97.5	93.3 (7)
SE derived from reported 95% CI, all 9 reports (incl. Park 2025 CI)	9	0.770	0.706–0.824	0.718–0.815	0.56–0.90	0.152	99.9	3082.0 (8)
Median-performing model per study instead of best model	8	0.737	0.645–0.813	0.663–0.800	0.43–0.91	0.250	98.4	142.6 (7)
Analysis domain not at high risk of bias (Lin, Chen, Bian)	3	0.788	0.407–0.953	0.635–0.888	0.00–1.00	0.423	95.2	23.6 (2)

Leave-one-out	k	Pooled AUC	HKSJ 95% CI	I², %
omit Lin 2011	7	0.756	0.689–0.813	96.7
omit Kang 2020	7	0.766	0.678–0.836	98.0
omit Li 2021	7	0.764	0.676–0.834	97.6
omit Yoshimura 2022	7	0.784	0.696–0.852	98.1
omit Chen 2025	7	0.782	0.691–0.851	98.1
omit Bian 2025	7	0.786	0.702–0.851	98.0
omit Katsin 2025	7	0.783	0.693–0.852	96.6
omit Sarullo 2025	7	0.787	0.703–0.852	94.7

Supplementary Table S7. Sensitivity and specificity: reconstructed 2×2 data, bivariate and univariate results

Study	Evaluation n	Events	Sens	Spec	Spec status	Threshold selection	TP	FN	TN	FP
Lin 2011	294	106	0.764	0.856	reported	Not reported	81	25	161	27
Kang 2020	222	126	0.8365	0.745	derived from precision 0.8116 + recall + prevalence	Not reported (default classifier output)	105	21	72	24
Lee 2020	282	151	0.7843	0.711	derived from precision 0.758 + recall + prevalence	Not reported	118	33	93	38
Li 2021	909	490	0.788	0.856	reported	Not reported	386	104	359	60
Yoshimura 2022	587	200	0.65	0.7	reported	Youden index	130	70	271	116
Park 2025	1101	672	0.688	0.714	reported	Fixed 0.5 cut-off on regressor output	462	210	306	123
Chen 2025	1622	276	0.7	0.68	derived from precision 0.31 + recall + prevalence	Youden index	193	83	915	431
Bian 2025	281	142	0.634	0.698	reported	Not reported	90	52	97	42

Bivariate analysis	k	Sensitivity (95% CI)	Specificity (95% CI)	τ^2 sens	τ^2 spec	ρ	LR+	LR−	DOR (95% CI)
Primary: independent cohorts (Lin, Lee, Li, Yoshimura, Chen, Bian)	6	0.723 (0.668–0.772)	0.758 (0.679–0.822)	0.079	0.216	0.93	2.99	0.37	8.16 (4.41–15.10)
Directly reported sensitivity AND specificity only (Lin, Li, Yoshimura, Park, Bian)	5	0.715 (0.649–0.773)	0.771 (0.683–0.840)	0.103	0.235	1.0	3.12	0.37	8.44 (4.07–17.50)
Alternative overlapping reports (Kang instead of Lee; Park instead of Chen)	6	0.726 (0.668–0.778)	0.773 (0.703–0.830)	0.099	0.181	1.0	3.2	0.35	9.03 (4.85–16.80)

Univariate pooling (secondary analysis)	k	Studies	Pooled	HKSJ 95% CI	95% PI	I², %
sensitivity	7	Lin 2011; Lee 2020; Li 2021; Yoshimura 2022; Chen 2025; Bian 2025; Katsin 2025	0.728	0.668–0.781	0.555–0.852	80.8
specificity	6	Lin 2011; Lee 2020; Li 2021; Yoshimura 2022; Chen 2025; Bian 2025	0.757	0.651–0.839	0.434–0.927	93.1

ρ estimated at or near its boundary (+1) with five or six studies; it indicates that studies with higher sensitivity also had higher specificity (differences in discrimination rather than in threshold) and is poorly determined. A summary ROC curve was therefore not drawn.

Supplementary Table S8. Predictive values and natural frequencies at the bivariate summary sensitivity (0.72) and specificity (0.76)

PIH prevalence	PPV	NPV	Flagged high risk /1,000	True positives /1,000	False positives /1,000	Missed PIH /1,000
10%	0.249	0.961	290	72	218	28
20%	0.427	0.916	338	145	194	55
30%	0.561	0.864	386	217	169	83
50%	0.749	0.732	482	361	121	139
60%	0.817	0.646	531	434	97	166

Supplementary Table S9. Exploratory subgroup analyses of the AUC (mixed-effects meta-regression, Knapp-Hartung F test)

Moderator	Level	k	Studies	Pooled AUC	HKSJ 95% CI	I^2, %	F (df)	p
Validation type	Temporal validation	2	Lin 2011; Bian 2025	0.812	not informative (k ≤ 2)	95.6	0.33 (4,3)	0.841
Validation type	Repeated cross-validation	1	Kang 2020	0.842	not informative (k ≤ 2)		0.33 (4,3)	0.841
Validation type	Held-out random split (test set)	3	Li 2021; Yoshimura 2022; Chen 2025	0.773	0.56–0.90	93.5	0.33 (4,3)	0.841
Validation type	Cross-validation / held-out (ambiguous)	1	Katsin 2025	0.732	not informative (k ≤ 2)		0.33 (4,3)	0.841
Validation type	5-fold cross-validation	1	Sarullo 2025	0.697	not informative (k ≤ 2)		0.33 (4,3)	0.841
Model family	Neural network	2	Lin 2011; Yoshimura 2022	0.82	not informative (k ≤ 2)	95.3	0.59 (2,5)	0.587
Model family	Tree-based	4	Kang 2020; Li 2021; Katsin 2025; Sarullo 2025	0.782	0.64–0.88	98.5	0.59 (2,5)	0.587
Model family	Regression-based	2	Chen 2025; Bian 2025	0.725	not informative (k ≤ 2)	32.7	0.59 (2,5)	0.587
Predictor timing tier	2 Pre-induction + induction drug doses	4	Lin 2011; Li 2021; Katsin 2025; Sarullo 2025	0.8	0.61–0.91	99.1	1.18 (2,5)	0.379
Predictor timing tier	3 Includes intra-induction haemodynamics	1	Kang 2020	0.842	not informative (k ≤ 2)		1.18 (2,5)	0.379
Predictor timing tier	1 Pre-induction data only	3	Yoshimura 2022; Chen 2025; Bian 2025	0.725	0.67–0.78	0.0	1.18 (2,5)	0.379
PIH threshold	SBP <90 mmHg-based composite	3	Lin 2011; Kang 2020; Li 2021	0.854	0.77–0.91	32.0	69.13 (2,5)	<0.001
PIH threshold	MAP <55 mmHg	3	Yoshimura 2022; Chen 2025; Katsin 2025	0.732	0.70–0.77	0.0	69.13 (2,5)	<0.001
PIH threshold	MAP <=65 mmHg (+/- relative fall)	2	Bian 2025; Sarullo 2025	0.697	not informative (k ≤ 2)	0.0	69.13 (2,5)	<0.001

Supplementary Table S10. Within-study comparisons of the best model with the other models evaluated on the same data

Study	Models	Best model	Best AUC	Median AUC	Range	Best − median	Linear comparator	Comparator AUC	Best − comparator
Chen 2025	4	Logistic regression	0.74	0.72	0.710–0.740	0.02	Logistic regression	0.74	0.0
Kang 2020	4	Random forest (feature set C)	0.8423	0.769	0.756–0.842	0.073	Logistic regression (feature set B)	0.7556	0.087
Katsin 2025	2	Random forest (AutoML)	0.732	0.686	0.639–0.732	0.046	Logistic regression	0.639	0.093
Lin 2011	2	ANN	0.893	0.866	0.840–0.893	0.026	Logistic regression (stepwise)	0.84	0.053
Park 2025	6	XGBoost regressor	0.772	0.768	0.685–0.772	0.004	ARD regression	0.685	0.087
Sarullo 2025	3	XGBoost	0.697	0.694	0.671–0.697	0.003	none		
Yoshimura 2022	4	Deep neural network	0.72	0.56	0.540–0.720	0.16	Linear discriminant analysis	0.56	0.16

Supplementary Methods. Formulas and derivations

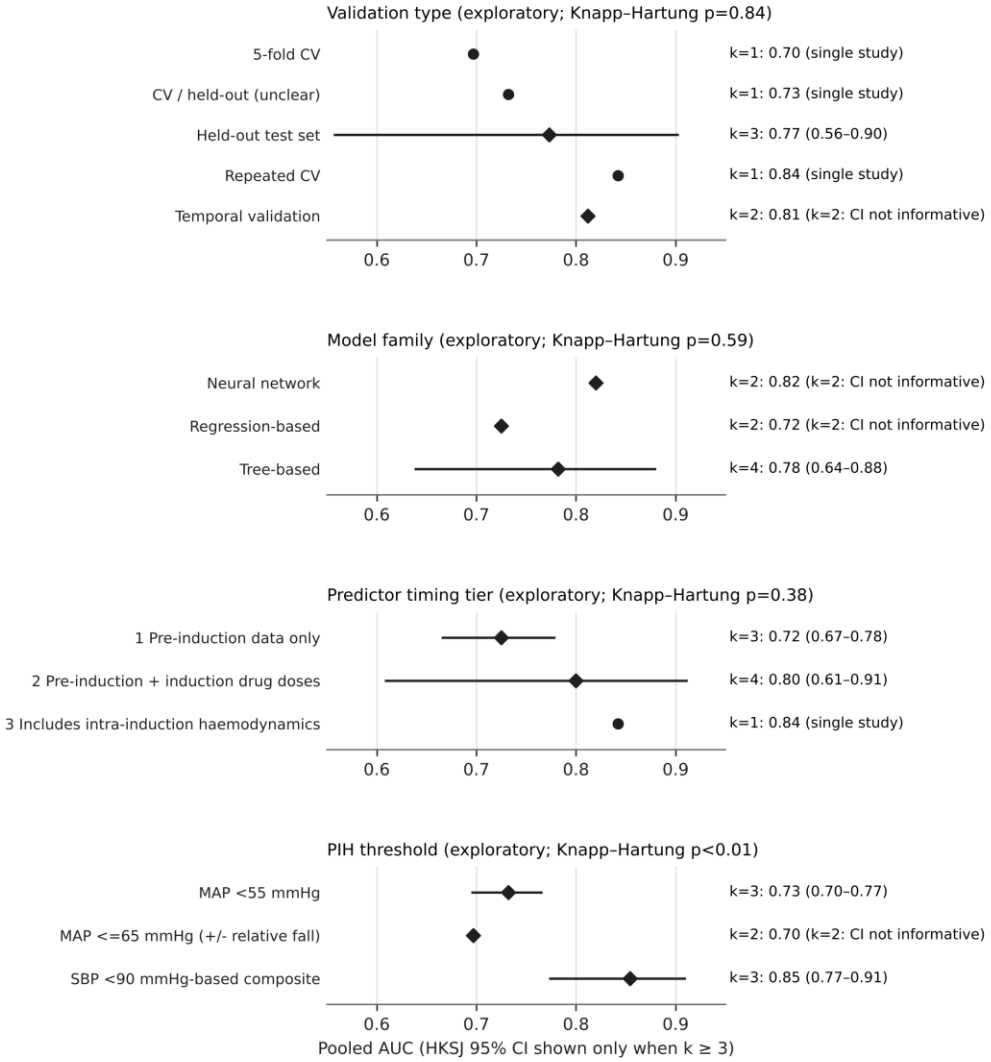
Hanley-McNeil standard error of an AUC (A) with n1 events and n0 non-events: $SE^2 = [A(1-A) + (n1-1)(Q1-A^2) + (n0-1)(Q2-A^2)]/(n1 \cdot n0)$, where $Q1 = A/(2-A)$ and $Q2 = 2A^2/(1+A)$. Logit-scale variance by the delta method: $Var[\text{logit}(A)] = SE^2/[A(1-A)]^2$.

Derived specificity from precision (PPV), recall (sensitivity) and prevalence: TP = sens × n1; FP = TP × (1 − PPV)/PPV; specificity = 1 − FP/n0. Accuracy recomputed from the derived 2×2 table was compared with the reported accuracy as a consistency check (Kang: 0.797 vs 0.795; Lee: 0.750 vs 0.749; Chen: 0.68 vs 0.67). Katsin 2025 failed this check (implied accuracy ≈0.60 vs reported 0.771), so specificity was not derived.

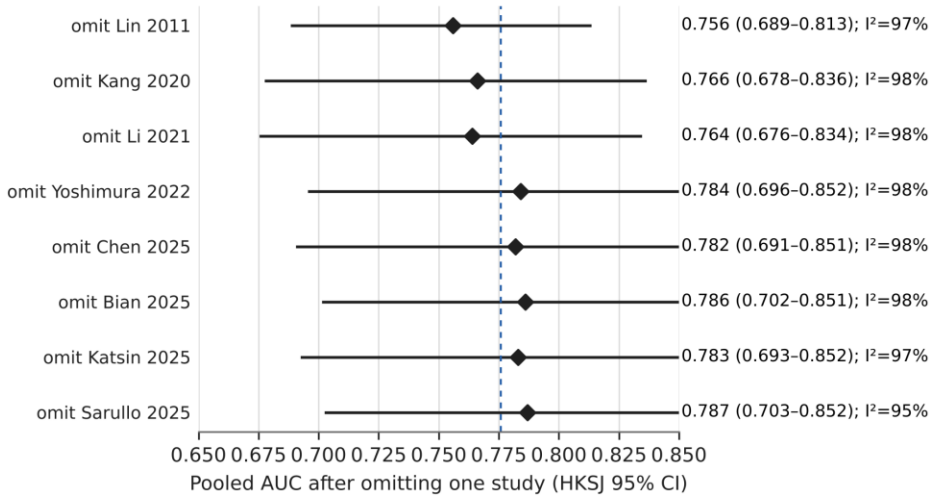
Random-effects model: REML estimate of τ^2; Hartung-Knapp-Sidik-Jonkman variance with the ad hoc modification that the HKSJ interval is never narrower than the Wald interval; t distribution with k−1 df. 95% prediction interval: $\mu \pm t(k-2) \cdot \sqrt{(\tau^2 + SE^2)}$.

Bivariate model (Reitsma 2005): (logit sens_i, logit spec_i) ~ N(μ, Σ + S_i), S_i = diag(1/TP+1/FN, 1/TN+1/FP); REML; 95% prediction region uses Σ + Cov(μ̂) with an F(2, k−2)-based radius.

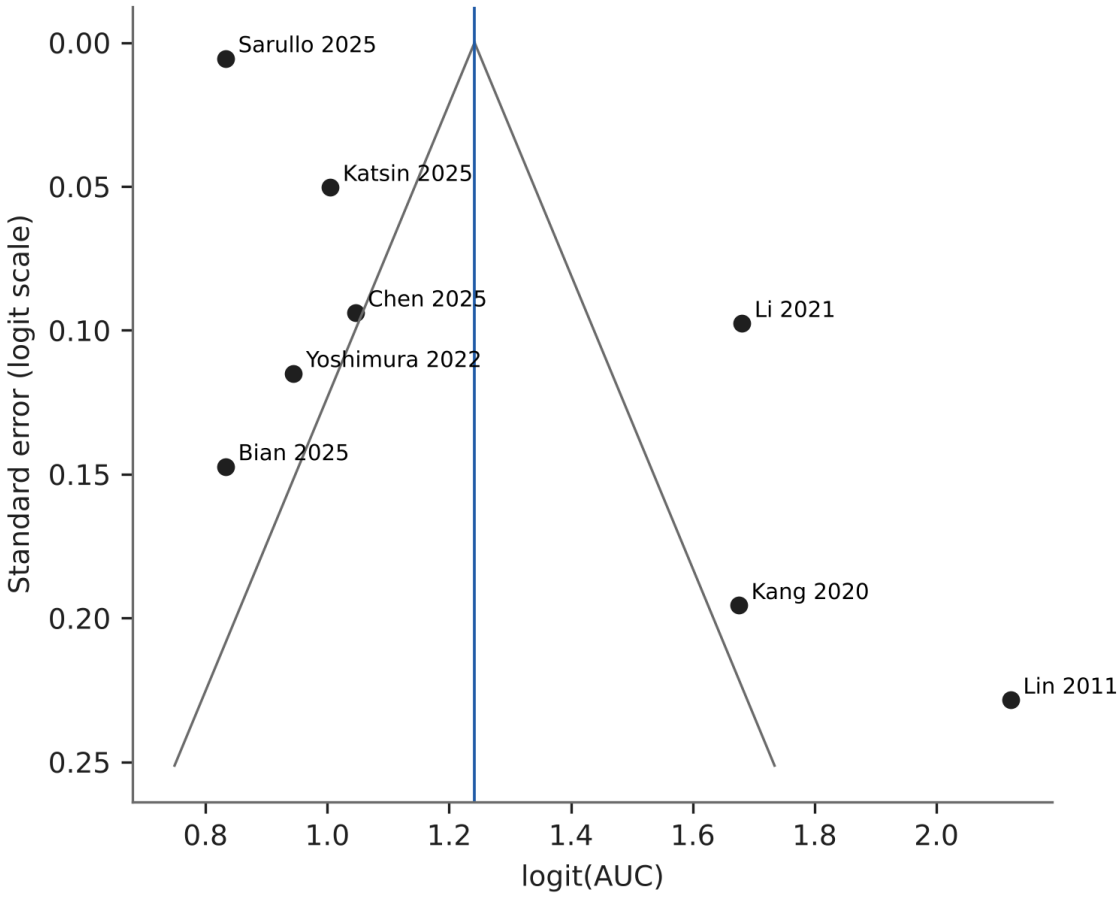
Software: Python 3.13 with NumPy, SciPy, pandas and matplotlib (scripts in Supplementary Data). The REML routine reproduces the metafor benchmark (dat.bcg: μ = −0.7145, τ^2 = 0.3132, I^2 = 92.22%); the bivariate routine recovered simulated parameters. An R script (metafor and mada) reproducing the main analyses is supplied for independent verification.



Supplementary Figure S1. Exploratory subgroup analyses of the pooled AUC by validation type, model family, predictor-timing tier and PIH definition (mixed-effects meta-regression; HKSJ 95% CIs shown only when k ≥ 3).



Supplementary Figure S2. Leave-one-out analysis of the pooled AUC (dashed line, primary estimate).



Supplementary Figure S3. Funnel plot of logit(AUC) against its standard error, shown descriptively; formal tests for small-study effects were not performed because fewer than ten studies were available.