Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and Performance Comparison of Machine Learning Models for Predicting Difficult Airway Based on Clinical Physical Measurements and Airway Ultrasound Parameters

Jinghao Sun
Second Affiliated Hospital of Zhengzhou University
Li Xia
sahxl@zzu.edu.cn

Second Affiliated Hospital of Zhengzhou University
Tianhao Sun
Bridger Investment Partners

Research Article

Keywords: Difficult airway, Machine learning, Airway ultrasound, General anesthesia

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11402308/v1

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

Additional Declarations: No competing interests reported.

Title Page

[Type of contribution]:

Research Article

[Title]:

Development and Performance Comparison of Machine Learning Models for Predicting Difficult Airway Based on Clinical Physical Measurements and Airway Ultrasound Parameters

[Author Names]:

Jinghao Sun[1], Master

Li Xia*[1], M.D.

Tianhao Sun[2], Master

1. Department of Anesthesiology and Perioperative Medicine, The Second Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan 450003, China

2. Data and Research Department, Bridger Investment Partners, Wuhan, Hubei 430061, China

[Version date]:

2026-10-13

[*Corresponding author]:

Name: Li Xia

Address: The Second Affiliated Hospital of Zhengzhou University, Jingba Road, Jinshui District, Zhengzhou, Henan 450003, China

E-mail addresses: sahxl@zzu.edu.cn

Development and Performance Comparison of Machine Learning Models for Predicting Difficult Airway Based on Clinical Physical Measurements and Airway Ultrasound Parameters

Keywords: Difficult airway, Machine learning, Airway ultrasound, General anesthesia

Abstract

Background: Difficult airway is an important cause of serious perioperative adverse events. Conventional preoperative assessment is simple but partly operator dependent and has limited predictive accuracy. We developed machine learning models combining clinical measurements and airway ultrasound parameters to support preoperative difficult airway assessment.

Methods: Seven predictors were collected from 1,000 adults undergoing elective surgery under general anesthesia: body mass index, neck circumference, mouth opening, sternomental distance, skin-to-epiglottis distance, tongue thickness, and hyomental distance ratio. Data were split into training (n=800) and test (n=200) sets. Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost), generalized additive model (GAM), and Naive Bayes were tuned by stratified 5-fold cross-validation using ROC-AUC. The principal metrics were the area under the receiver operating characteristic curve (ROC-AUC), average precision (AP), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), balanced accuracy, and F1 score.

Results: Mean cross-validation ROC-AUC values were 0.818 for Random Forest, 0.797 for XGBoost, 0.790 for GAM, 0.786 for Logistic Regression, and 0.770 for Naive Bayes. The test set included 24 difficult airway cases. Test-set ROC-AUC values were 0.802 for GAM, 0.783 for XGBoost, 0.777 for Random Forest, and 0.771 for Logistic Regression. At the 0.5 threshold, Logistic Regression had the highest sensitivity (0.542) and a negative predictive value of 0.930; XGBoost had sensitivity 0.458, specificity 0.892, and F1 score 0.407.

Conclusions: Models combining clinical and airway ultrasound measurements showed some discrimination for difficult airway prediction. Given its sensitivity, negative predictive value, and interpretability, Logistic Regression is a candidate for calibration, threshold optimization, and external validation.

1. Introduction

Difficult airway is a major anesthetic challenge and contributor to serious perioperative adverse events. An unanticipated difficult airway may cause difficult mask ventilation, failed tracheal intubation, hypoxemia, aspiration, or life-threatening complications[1,2]. Its management requires expertise and complex interventions. Accurate preoperative prediction is therefore important for planning airway management and protecting perioperative patient safety[3,4].

Common preoperative assessments include the Mallampati classification, mouth opening, thyromental distance, neck circumference, and body mass index (BMI), which are widely used because they are simple to perform[3,5]. However, conventional airway

assessment often depends on operator experience and is subject to some degree of subjectivity[6]. Moreover, the predictive performance of any single measure is usually limited and cannot fully represent the complexity of a patient's airway anatomy[7,8]. Improving the accuracy of preoperative difficult airway prediction therefore remains a major concern in anesthesiology.

In recent years, ultrasound has increasingly been used for preoperative difficult airway assessment because it is noninvasive, portable, and repeatable[7,9]. Many patients at risk of difficult intubation have anatomical airway abnormalities. Although routine physical examination cannot directly visualize some deeper airway structures, ultrasound can display airway anatomy and provide objective quantitative measurements, offering an additional source of information for difficult airway prediction[10,11]. With advances in computational methods, artificial intelligence, particularly machine learning, has been used increasingly in clinical prediction models[12]. These methods can integrate multiple variables and capture potential nonlinear relationships and interactions, and may improve the accuracy of preoperative difficult airway prediction. To date, conventional airway assessments, ultrasound parameters, and artificial intelligence models have all been investigated for difficult airway prediction and management[13]. Existing studies, however, remain limited by a single source of variables, limited model generalizability, or inadequate clinical practicality[14]. Studies that combine conventional physical examination measures with airway ultrasound parameters and then develop machine learning prediction models remain relatively scarce. In addition, most existing studies focus only on difficult intubation and overlook the broader definition of difficult airway: during airway management, the occurrence of difficult facemask ventilation,

difficult laryngoscopy, difficult supraglottic airway ventilation, difficult or failed tracheal intubation, difficult or failed tracheal extubation, difficult or failed invasive airway, or inadequate ventilation may be classified as a difficult airway[3].

Given the limitations of current difficult airway assessment methods and machine learning prediction models, a practical and accurate model suitable for routine clinical use is still needed. Accordingly, this study developed machine learning models for difficult airway prediction using preoperative clinical assessment variables and airway ultrasound parameters and evaluated their predictive performance, with the aim of providing a reference for preoperative difficult airway assessment and individualized airway management.

2. Methods

2.1 Study Design and Participants

This prospective observational study enrolled 1,000 adult patients undergoing elective surgery under general anesthesia at the Second Affiliated Hospital of Zhengzhou University. The study protocol was reviewed and approved by the hospital's ethics committee on June 8, 2026 (approval no. KY2026232). Written informed consent was obtained from all eligible patients, and all participants consented to publication of their images, clinical data, and other relevant information. The study was registered with the Chinese Clinical Trial Registry, a primary registry of the World Health Organization International Clinical Trials Registry Platform (registration no. ChiCTR2600127067; website: http://www.chictr.org.cn). The inclusion criteria were: (1) age 18 years or older; (2) scheduled elective surgery under general

anesthesia; (3) American Society of Anesthesiologists (ASA) Physical Status I-III; and (4) voluntary provision of written informed consent. The exclusion criteria were: (1) central nervous system disease or psychiatric illness; (2) otolaryngologic disease or a related surgical history that might affect airway structure, or a known difficult airway requiring fiberoptic bronchoscope-guided intubation; and (3) severe maxillofacial deformity or cervical disease that might affect airway assessment. Because this study developed a binary prediction model for difficult airway, the sample size was estimated initially using the population-proportion method. Under the normal approximation to the sampling distribution of a sample proportion in a large sample, a 90% confidence interval, a 3% margin of error, and the most conservative proportion estimate ($p=0.5$) yielded a required sample size of approximately 752 patients. Allowing for study implementation needs and potential data loss, the planned enrollment was at least 800 patients; 1,000 patients were ultimately enrolled. Data were collected from 1,000 patients between July and September 2026. Among them, 344 were male and 656 were female, and the mean age was 53.5 years. ASA Physical Status ranged from I to III: 19.2% were ASA I, 63.4% were ASA II, and 17.4% were ASA III.

2.2 Clinical Airway Assessment and Ultrasound Measurements

During the preoperative visit, anesthesiologists collected seven prespecified airway-related predictors using standardized procedures. Of 1,000 patients, 800 formed the training set for model fitting, 5-fold cross-validation, and hyperparameter selection; 200 were reserved exclusively for final testing. Study airway assessment results were concealed from the

anesthesiologist providing anesthesia, who could perform a routine preoperative assessment. On arrival in the operating room, routine monitoring of baseline vital signs was initiated, including electrocardiography (ECG), noninvasive blood pressure (NBP), oxygen saturation (SpO2), and end-tidal carbon dioxide (EtCO2). Anesthesia was induced intravenously with remimazolam (0.2-0.3 mg/kg), sufentanil (0.5-1.0 μg/kg), cisatracurium (0.15-0.2 mg/kg), and etomidate (0.3-0.6 mg/kg). Patients were adequately preoxygenated before induction, followed by mask ventilation and assessment of its difficulty. After complete neuromuscular relaxation, an anesthesiologist with more than five years of clinical experience inserted a supraglottic airway device or performed videolaryngoscopy intubation, assessed the laryngeal view, and evaluated difficult supraglottic airway ventilation, difficult laryngoscopy, difficult or failed tracheal intubation, inadequate ventilation, or other difficult airway events. Following the basic principles of the ASA difficult airway guidelines, difficult airway was defined as difficulty or failure in establishing or maintaining an effective airway[3]. The outcome was positive if airway management involved difficult facemask ventilation, difficult laryngoscopy, difficult supraglottic airway ventilation, difficult or failed tracheal intubation, difficult or failed tracheal extubation, difficult or failed invasive airway, or inadequate ventilation. Patients with none of these events were classified as having a non-difficult airway. When a difficult airway was encountered, management was undertaken promptly in accordance with the relevant guidelines, with assistance from a senior physician when necessary. For patients with an anticipated difficult airway, the anesthesiologist developed an airway management plan based on the specific risks. Patients known to require fiberoptic bronchoscope-guided intubation were excluded as prespecified.

The seven prespecified airway-related predictors were body mass index (BMI), neck circumference, mouth opening, sternomental distance (SMD), skin-to-epiglottis distance (SED), tongue thickness (TT), and hyomental distance ratio (HMDR), measured as follows:

(1) Body mass index (BMI): calculated as weight (kg) divided by height squared (m^2).

(2) Neck circumference: measured with a flexible tape at the cricoid cartilage level with the head and neck neutral.

(3) Mouth opening: the distance between the upper and lower incisors maximal mouth opening.

(4) Sternomental distance (SMD): the distance from the mentum to the suprasternal notch with the head fully extended[13].

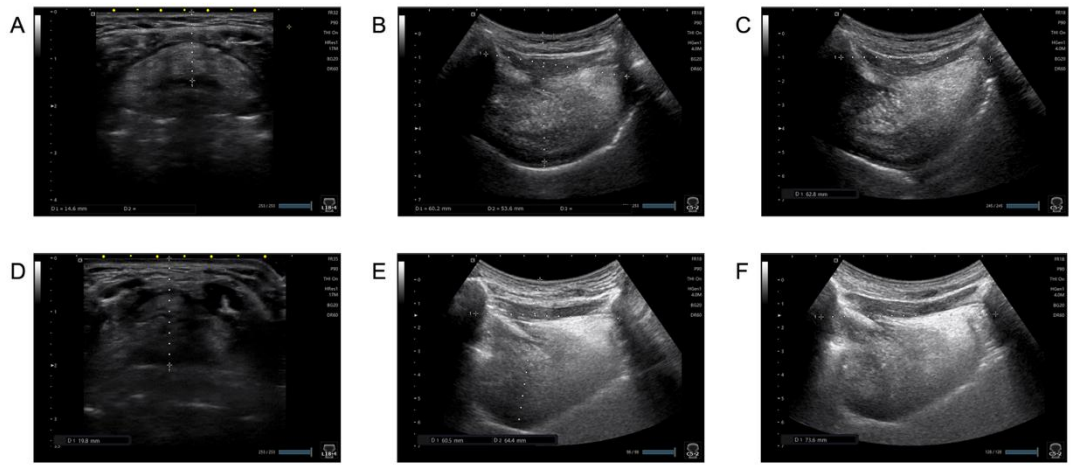
(5) Skin-to-epiglottis distance (SED): with the patient supine, a linear-array transducer was placed transversely just above the thyroid cartilage. The skin-to-epiglottis distance was measured after the epiglottis and thyroid cartilage were clearly visualized[15,16]. Representative ultrasound images are shown in Figure 1A and 1D.

(6) Tongue thickness (TT): a low-frequency transducer was positioned in the midline at the head-neck junction and oriented sagittally. The tongue was kept closely apposed to the mandible to reduce interference from air between tissues. After visualizing the tongue from tip to base, the distance from the skin beneath the mandible to the tongue surface was measured at its thickest part[16,17]. Representative ultrasound images are shown in Figure 1B and 1E.

(7) Hyomental distance ratio (HMDR): the patient lay supine without a pillow with the head and neck examined in neutral and extended positions. A high-frequency linear-array transducer located the hyoid bone. The distance from the inferior mandibular border to the hyoid bone was measured in each position. HMDR = hyomental distance in the extended position / hyomental distance in the neutral position. The ratio is dimensionless[9,13]. Representative ultrasound images are shown in Figure 1B, 1C, 1E, and 1F.

Figure 1

Comparison of ultrasound parameter images between patients with non-difficult and difficult airways.



(A)-(C) Images from a patient with a non-difficult airway: (A) skin-to-epiglottis distance, 1.46 cm; (B) hyomental distance in the neutral position, 6.02 cm, and tongue thickness, 5.36 cm; (C) hyomental distance in the extended position, 6.28 cm. (D)-(F) Images from a patient with a difficult airway: (D) skin-to-epiglottis distance, 1.98 cm; (E) hyomental distance in the

186 neutral position, 6.05 cm, and tongue thickness, 6.44 cm; (F) hyomental distance in the
187 extended position, 7.36 cm.

188

189 2.3 Data Preprocessing and Model Development

190 After basic cleaning, difficult airway was encoded as a binary outcome. Patient number,
191 name, surgery date, and type of difficult airway were excluded from the models. Height,
192 weight, and the two source hyomental distances were not included alongside BMI and
193 HMDR, avoiding redundant use of derived variables and their components.

194 Logistic Regression, GAM, and Naive Bayes included standardization in the training
195 pipeline, and Random Forest and XGBoost did not require it. All imputation and
196 standardization remained within the training pipeline, and the test set was used only for final
197 evaluation to prevent leakage into training[18].

198 Within the training set, 5-fold cross-validation reduced dependence on a single split and
199 supported model and hyperparameter selection. In each fold, approximately 80% of the
200 training data were used for fitting and 20% served for validation. Grid-search cross-validation
201 (GridSearchCV) evaluated prespecified parameter grids, and selected parameter combinations
202 by mean ROC-AUC across 5-fold cross-validation[19]. Imputation, standardization, class-
203 weight specification, and model fitting used only each training fold, and candidate parameters
204 were evaluated on its validation fold. Selected models were refitted on all 800 patients and
205 evaluated on the 200 test patients. All five models used the same seven predictors.

206 To address the low prevalence of difficult airway, Logistic Regression used

class_weight= “balanced”, which adjusts weights according to class frequencies in scikit-learn. Random Forest used class_weight= “balanced_subsample”, and XGBoost used scale_pos_weight. GAM and Naive Bayes were unweighted. Classification results from the different models should be interpreted alongside training weights, probability calibration, and decision thresholds. No oversampling, including the synthetic minority oversampling technique (SMOTE), was used.

Of the full dataset, 800 patients were assigned to the training set, in which stratified 5-fold cross-validation was used for hyperparameter selection and control of overfitting risk. Logistic Regression used the regularization parameter C=0.1. Random Forest used 500 base estimators (n_estimators=500), a minimum of 3 samples per leaf (min_samples_leaf=3), and a maximum tree depth of 4 (max_depth=4). XGBoost used a learning rate of 0.03 (learning_rate=0.03), a maximum tree depth of 2 (max_depth=2), and 200 estimators (n_estimators=200). GAM used the smoothing-penalty parameter lam=3.0.

2.3.1Data

Before model training, descriptive statistics summarized the seven predictors in the training set (Additional file 1: Table S1). Counts, means, standard deviations, and quartiles were examined for completeness, plausible distributions, extreme values, and derived-variable abnormalities, to avoid interpreting data-quality problems as clinical risk signals.

2.3.2 Models

Interpretable models (Logistic Regression, GAM, and Naive Bayes) were compared with black-box models. Their continuous predictors were standardized to a common scale to

reduce unit-related effects on scale-sensitive fitting. Logistic Regression provides good interpretability, and its linear predictor describes the relationship between the variables and the log-odds of the outcome. GAM explores nonlinear risk relationships, and Naive Bayes provides a simple probabilistic comparator. Random Forest and XGBoost respectively represent bagging and boosting. Bagging fits learners to repeated bootstrap samples and aggregates predictions to reduce variance and overfitting risk. Boosting sequentially adds learners to optimize the ensemble loss, combining simple models into a stronger predictor. Both tree models assessed complex nonlinearities and interactions, and XGBoost was included as a candidate based on overall classification performance.

GAM, Random Forest, and XGBoost can all be used to assess nonlinear relationships, with GAM retaining some interpretability. Because further evaluation of complex nonlinear relationships in the data was considered necessary, a supplementary Logistic Regression analysis using restricted cubic splines was conducted within the training set[20]. For each continuous predictor, linear and nested spline Logistic Regression models fitted to the same sample were compared using a likelihood-ratio test; the fitting approach, test degrees of freedom, Akaike information criterion (AIC), and Bayesian information criterion (BIC) were reported. Multivariable models with all predictors entered linearly or as spline were also compared using training-set internal 5-fold cross-validation ROC-AUC and average precision (AP).

2.4. Model Performance Evaluation and Statistical Analysis

Classification metrics were accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), balanced accuracy, and F1 score, defined below:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Sen = \frac{TP}{TP+FN} \quad (2)$$

$$Spe = \frac{TN}{FP+TN} \quad (3)$$

$$PPV = \frac{TP}{TP+FP} \quad (4)$$

$$NPV = \frac{TN}{TN+FN} \quad (5)$$

$$Balanced\ accuracy = \frac{Sen+Spe}{2} \quad (6)$$

$$F1 = \frac{2TP}{2TP+FP+FN} \quad (7)$$

TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The result in Equation (6) was calculated from the results of Equations (2) and (3).

To evaluate comprehensively the discrimination of the difficult airway prediction models, ROC-AUC and AP were analyzed. ROC-AUC ranges from 0 to 1, with 0.5 indicating random ranking, and values closer to 1 indicate better discrimination[21,22]. Under class imbalance, AP summarizes the positive-class precision–recall trade-off and is useful for evaluating difficult airway prediction[22]. AP weights precision by increments in recall and differs from trapezoidal precision-recall area under the curve (PR-AUC).

Because low prevalence can inflate apparent performance based on accuracy alone, all discrimination and classification metrics were considered jointly. Test-set 95% confidence

intervals (CIs) were calculated from 2,000 stratified nonparametric bootstrap resamples.

Data processing, model development, and graphical analyses used macOS 26.4.1 and Python 3.13.12 in a conda environment, with pandas 3.0.1, NumPy 2.4.3, scikit-learn 1.8.0, XGBoost 3.2.0, pyGAM 0.12.0, and Matplotlib 3.10.8. Test-set outcomes and prediction scores generated the ROC and precision–recall curves. The matplotlib.pyplot module plotted feature interpretations from the final models' regression coefficients and the importance measures used.

3. Results

During the study period, 1,000 patients underwent preoperative clinical assessment and airway ultrasound assessment. According to the basic principles of the American Society of Anesthesiologists guidelines for difficult airway management, 873 patients had a non-difficult airway and 127 had a difficult airway. Table 1 summarizes clinical characteristics and airway-related predictors. Compared with patients with a non-difficult airway, those with a difficult airway were slightly younger, with similar sex distributions. They had higher BMI, larger neck circumference, slightly smaller mouth opening, shorter sternomental distance, greater skin-to-epiglottis distance and tongue thickness, and slightly lower HMDR. These descriptive differences suggest differing distributions in this sample. Table S2 in Additional file 1 shows the specific difficult airway types among the 127 affected patients: difficult facemask ventilation in 62, difficult laryngoscopy in 43, difficult supraglottic airway ventilation in 19, inadequate ventilation in 2, and difficult or failed tracheal extubation in 1.

Table 1

Basic Patient Characteristics and Airway-Related Assessment Results.

	All	Non-difficult airway	Difficult airway
	n=1000	n=873	n=127
Age, years	55(42-64)	55(43-64)	51(38-63)
Sex, female	656(65.60%)	574(65.75%)	82(64.57%)
BMI, kg/m^2	24.32(21.22-27.49)	24.02(21.10-27.28)	26.18(22.33-30.08)
Neck circumference, cm	37.70(35.90-39.80)	37.50(35.70-39.70)	39.80(37.30-43.80)
Mouth opening, cm	4.10(3.80-4.50)	4.20(3.80-4.50)	4.10(3.80-4.50)
Sternomental distance, cm	13.70(12.80-14.70)	13.70(12.90-14.70)	13.40(11.65-14.30)
Skin-to-epiglottis distance, cm	1.84(1.68-2.18)	1.83(1.67-2.18)	2.27(1.83-2.51)
Tongue thickness, cm	5.97(5.39-6.59)	5.92(5.37-6.52)	6.54(5.84-6.89)
Hyomental distance ratio	1.10(1.06-1.14)	1.10(1.06-1.14)	1.09(1.06-1.12)

Categorical variables are presented as n (%); non-normally distributed continuous variables as median (first-third quartile).

Abbreviations: BMI, body mass index.

Internal 5-fold cross-validation in the training set yielded mean ROC-AUC values of 0.77-0.82 (Additional file 1: Table S3). Random Forest had the highest mean ROC-AUC, 0.818±0.077; XGBoost, GAM, Logistic Regression, and Naive Bayes yielded 0.797±0.083, 0.790±0.092, 0.786±0.075, and 0.770±0.083, respectively. Random Forest also had the

highest AP, 0.439±0.098. At the default threshold, models showed different sensitivity-specificity trade-offs under low prevalence. Logistic Regression had the highest sensitivity (0.719±0.131), but relatively low specificity (0.729±0.068), whereas GAM had the highest specificity (0.968±0.019) but low sensitivity (0.184±0.020). XGBoost had the highest balanced accuracy (0.754±0.067). F1 scores were similar for Random Forest, XGBoost, Naive Bayes, and Logistic Regression and lower for GAM.

Test-set ROC-AUC values were 0.802, 0.783, 0.777, 0.771, and 0.716 for GAM, XGBoost, Random Forest, Logistic Regression, and Naive Bayes, respectively (Table 2). GAM had the highest AP, at 0.322, while values for XGBoost, Random Forest, Logistic Regression, and Naive Bayes were 0.292, 0.289, 0.286, and 0.242, respectively. Although GAM and XGBoost had relatively high ROC-AUC values, at the default threshold of 0.5 GAM had low sensitivity and XGBoost had intermediate sensitivity. In comparison, Logistic Regression had the highest sensitivity, at 0.542, with a specificity of 0.830 and balanced accuracy of 0.686. XGBoost had a sensitivity of 0.458, a specificity of 0.892, balanced accuracy of 0.675, and a relatively high F1 score of 0.407. Random Forest had a specificity of 0.920 and an F1 score of 0.311. Naive Bayes had a sensitivity of 0.125, a specificity of 0.926, balanced accuracy of 0.526, and an F1 score of 0.150. The models showed some discrimination, but their ability to identify positive cases at the default threshold was limited. Thresholds should therefore be optimized further according to the clinical screening objective.

The test set comprised 200 patients, including 24 with a difficult airway and 176 with a non-difficult airway. The positive-class prevalence was 12.0%, close to the 12.7% prevalence in the full sample. Table 3 shows the confusion matrices at threshold of 0.5. True

positives/false negatives were 13/11 for Logistic Regression, 7/17 for Random Forest, 11/13 for XGBoost, 1/23 for GAM, and 3/21 for Naive Bayes. Corresponding false positives/true negatives were 30/146, 14/162, 19/157, 6/170, and 13/163. These counts show marked sensitivity-specificity trade-offs, with Logistic Regression relatively sensitive at threshold 0.5.

Table 2

Predictive Performance of the Models in the Test Set.

Model	ROC-AUC	AP	Acc	Sen	Spe	PPV	NPV	BAcc	F1 score
GAM	0.802(0.706-0.882)	0.322(0.235-0.501)	0.855(0.830-0.880)	0.042(0.000-0.125)	0.966(0.938-0.989)	0.143(0.000-0.500)	0.881(0.874-0.891)	0.504(0.472-0.551)	0.065(0.000-0.194)
XGBoost	0.783(0.694-0.860)	0.292(0.216-0.456)	0.840(0.790-0.885)	0.458(0.250-0.667)	0.892(0.846-0.938)	0.367(0.229-0.519)	0.924(0.899-0.950)	0.675(0.577-0.773)	0.407(0.254-0.556)
Random Forest	0.777(0.682-0.858)	0.289(0.214-0.447)	0.845(0.800-0.885)	0.292(0.125-0.458)	0.920(0.881-0.960)	0.333(0.167-0.533)	0.905(0.885-0.928)	0.606(0.520-0.704)	0.311(0.143-0.480)
Logistic Regression	0.771(0.672-0.856)	0.286(0.200-0.423)	0.795(0.740-0.850)	0.542(0.333-0.750)	0.830(0.773-0.881)	0.302(0.204-0.421)	0.930(0.902-0.959)	0.686(0.581-0.787)	0.388(0.258-0.516)
Naive Bayes	0.716(0.596-0.825)	0.242(0.179-0.380)	0.830(0.790-0.865)	0.125(0.000-0.250)	0.926(0.886-0.966)	0.188(0.000-0.400)	0.886(0.872-0.903)	0.526(0.466-0.600)	0.150(0.000-0.308)

Acc, Sen, Spe, PPV, NPV, BAcc, and F1 score were calculated at the common classification threshold of 0.5. Metrics are presented as point estimates (95% CI); 95% CIs were calculated using 2,000 repetitions of stratified nonparametric bootstrap resampling. The test set included 200 patients, 24 of whom had a difficult airway.

Abbreviations: Acc: accuracy; AP: average precision; BAcc: balanced accuracy; NPV: negative predictive value; PPV: positive predictive value; ROC-AUC, the area under the receiver operating characteristic curve; Sen: sensitivity; Spe: specificity.

Table 3

Confusion Matrices for the Models at the Default Threshold of 0.5.

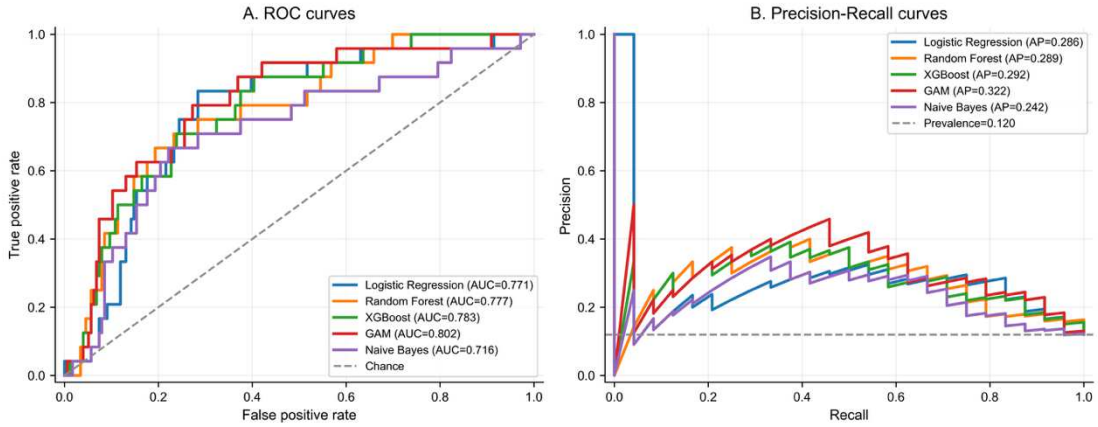
Model	TN	FP	FN	TP
Logistic Regression	146	30	11	13
Random Forest	162	14	17	7
XGBoost	157	19	13	11
GAM	170	6	23	1
Naive Bayes	163	13	21	3

Abbreviations: TN, true negatives; FP, false positives; FN, false negatives; TP, true positives.

Tables 2 and 3 show that ROC-AUC rankings did not fully reflect classification performance at the default threshold. GAM and XGBoost had relatively high ROC-AUC. XGBoost had the highest F1 score (0.407) but missed 13 cases. Logistic Regression identified 13 cases and was readily interpretable, whereas Random Forest had high specificity but identified only 7.

Figure 2

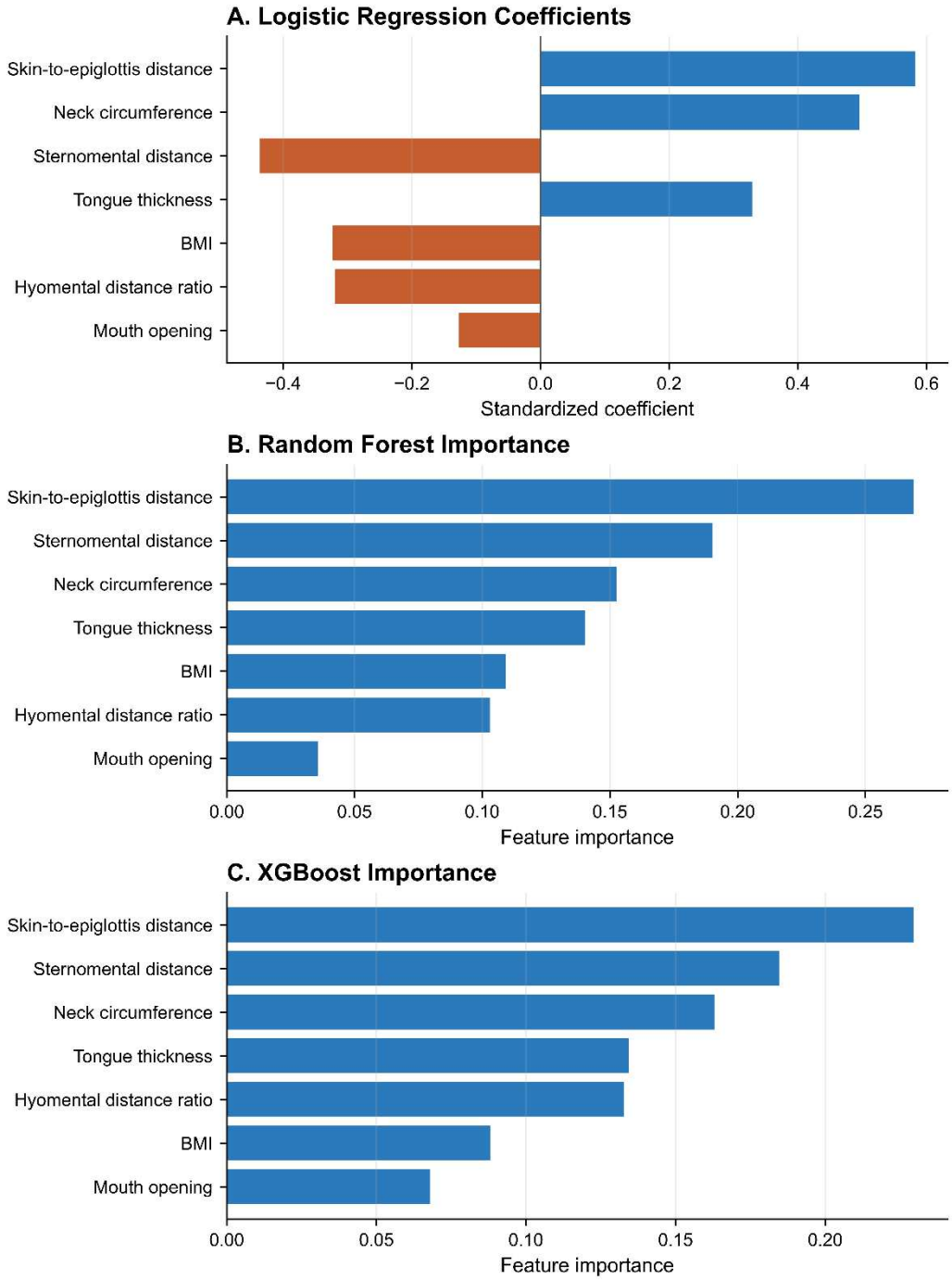
ROC and Precision-Recall Curves for the Models in the Test Set.



Panel A shows the ROC curves. Panel B shows the Precision-Recall curves (n=200, including 24 patients with a difficult airway). The dashed diagonal line in Panel A represents chance discrimination. The dashed horizontal line in Panel B marks the test-set difficult airway prevalence (12.0%).

Figure 3

Feature Effects and Importance Across Models.



Panel A shows standardized coefficients from Logistic Regression. Panels B and C show feature importance from Random Forest and XGBoost, respectively.

The ROC and Precision-Recall curves provide complementary views of test-set

discrimination (Figure 2). Every model had ROC-AUC >0.5. Under low positive-class prevalence, the PR curves give a more direct view of positive-class prediction and generally lay above the prevalence baseline. GAM had the highest AP, and all five models showed some ranking ability. Neither ROC-AUC nor AP alone establishes screening suitability at low prevalence; sensitivity, NPV, and the confusion matrix at the default threshold must also be considered.

Further effects and importance showed some consistency across models. The standardized Logistic Regression coefficients (Figure 3A) were positive for skin-to-epiglottis distance, neck circumference, and tongue thickness, and negative for sternomental distance, BMI, HMDR, and mouth opening, indicating corresponding associations with predicted difficult airway risk after multivariable adjustment. These coefficients describe within-model associations conditional on the other predictors and may be influenced by predictor correlations. The two tree-based ensembles, Random Forest and XGBoost, were also compared. In both tree models (Figures 3B and 3C), skin-to-epiglottis distance and sternomental distance had relatively high importance. In Random Forest, feature importance ranked as follows: skin-to-epiglottis distance, sternomental distance, neck circumference, tongue thickness, BMI, HMDR, and mouth opening. In XGBoost, skin-to-epiglottis distance ranked first, followed by sternomental distance, neck circumference, tongue thickness, HMDR, BMI, and mouth opening. Overall, airway ultrasound parameters, particularly skin-to-epiglottis distance and tongue thickness, together with sternomental distance and neck circumference among the conventional physical assessment measures, contributed relatively strongly to prediction.

Negative coefficients for mouth opening, sternomental distance, and HMDR are consistent with greater risk when working space or head and neck movement is less favorable. Considering the coefficient directions from Logistic Regression together with the importance values from the tree models, anterior neck soft-tissue thickness, tongue thickness, neck circumference, and measures related to head and neck movement may serve as principal interpretive clues, but these findings do not establish causal relationships with difficult airway.

Training-set restricted cubic spline Logistic Regression analyses (Additional file 1: Table S4) showed evidence of nonlinearity for skin-to-epiglottis distance (likelihood-ratio test [LRT] $P=0.007$, $\Delta AIC=-5.887$) and sternomental distance (LRT $P<0.001$, $\Delta AIC=-14.181$), but no clear evidence for BMI, neck circumference, mouth opening, tongue thickness, or HMDR. For the multivariable spline and linear models, respectively, AIC was 521.803 versus 522.312, BIC 624.865 versus 559.789, internal 5-fold cross-validation ROC-AUC 0.802±0.093 versus 0.783±0.073, and AP 0.421±0.117 versus 0.392±0.089. These results suggest that the spline model may modestly improve ranking, but its higher BIC indicates a greater complexity cost and warrants cautious interpretation.

Considering model performance, clinical screening needs, and interpretability, Logistic Regression was selected as a candidate model for subsequent validation. Other models had higher ROC-AUC, but no stable performance advantage was demonstrated. In addition, some models had insufficient sensitivity at the default threshold of 0.5, creating a risk of missing true difficult airway cases. By comparison, Logistic Regression has a simple structure and clear directions of variable effects, facilitating interpretation of the relationships between

variables and model-predicted risk; XGBoost and Random Forest may serve as candidate models for further performance comparison. The simplicity of Logistic Regression is also an important advantage. If models meet prespecified clinical requirements and added complexity provides no additional clinical benefit, a simple, interpretable model may be preferred[23,24].

Accordingly, Logistic Regression was therefore the principal candidate model for subsequent calibration, threshold optimization, and external validation, whereas the other machine learning models served as reference models for comparing predictive performance across algorithms.

4. Discussion

This study developed and evaluated machine learning models for difficult airway prediction using preoperative clinical physical measurements and airway ultrasound parameters. The results showed some discrimination and relatively favorable predictive performance across several metrics, suggesting that joint analysis of multiple parameters may aid preoperative risk identification.

Preoperative airway assessment should inform preparation for both ventilation and tracheal intubation. In our cohort, difficult facemask ventilation was the most frequent difficult airway event, followed by difficult laryngoscopy and difficult supraglottic airway ventilation. This pattern supports considering a broader airway outcome when studying perioperative risk. The ASA and Difficult Airway Society guidelines emphasize assessment and advance planning[3,4]. Earlier recognition of possible difficulty can inform preparation of

equipment, availability of experienced assistance, and alternative strategies for maintaining oxygenation.

The contributions of conventional measurements and ultrasound parameters are clinically relevant. Neck circumference and sternomental distance describe external anatomical dimensions, while skin-to-epiglottis distance and tongue thickness add information about internal soft tissues[9,11]. In the logistic regression model, greater skin-to-epiglottis distance, tongue thickness, and neck circumference were associated with higher predicted risk, whereas greater sternomental distance was associated with lower predicted risk. Considering these measurements together may help clinicians interpret the anatomical features underlying a risk estimate, rather than relying on a single bedside test.

Our findings are consistent with Demir Senoglu et al., who identified neck circumference, skin-to-epiglottis distance, and tongue thickness among the strongest predictors in a prospective cohort of 329 patients[13]. The agreement in anatomical predictors supports their relevance across studies, although that study focused on difficult intubation. Airway ultrasound offers a noninvasive means of obtaining measurements that are not directly available from physical examination[7,9]. Its potential role in preoperative assessment is therefore to complement established clinical measurements with additional anatomical information.

Among the evaluated models, logistic regression had the highest sensitivity at the default threshold and offered a readily interpretable relationship between measurements and predicted risk. This combination makes it a practical candidate for further clinical evaluation[23,24]. In

446 airway assessment, threshold selection should reflect the consequences of missed difficulty
447 and unnecessary alerts[22,25]. The intended role of a future risk estimate would be to support
448 the anesthesiologist's overall assessment, incorporating airway history and examination
449 findings, and to guide the level of preparation before induction.

450 Despite these findings, the study has several limitations. First, this single-center study
451 included adults undergoing elective surgery under general anesthesia and excluded some
452 patients with known difficulty or obvious anatomical airway abnormalities. The applicability
453 to emergency patients, pediatric patients, and the excluded high-risk populations remains
454 unclear. Differences in patient characteristics, anesthesia management processes, and the
455 operational difficult airway criteria may affect performance across centers, requiring
456 validation in other populations. Second, the limited sample, particularly only 24 difficult
457 airway cases in the test set, constrained performance assessment and estimate
458 precision[25,26]. These initial estimates may be unstable, and insufficient sample size may
459 increase the risk of overfitting[27]. In addition, physical and ultrasound measurements may
460 also vary with operator experience, measurement methods, and patient cooperation,
461 introducing potential bias. Finally, the present study focused mainly on internal performance
462 evaluation, and external validation and prospective validation of clinical application have not
463 been performed[21].

464 Future work may address these limitations in several ways. First, larger multicenter
465 validation could assess model robustness and applicability. Variable-selection strategies could
466 also be refined to develop simpler, more implementable prediction tools without
467 compromising model performance. Beyond conventional physical measurements and

ultrasound parameters, future studies could add anesthesia history[8,28], imaging features, or dynamic assessment measures to improve model performance. Finally, developing the model into a simple risk-scoring system or clinical decision-support tool may facilitate its use in preanesthetic assessment[29,30].

5. Conclusion

In conclusion, prediction models developed from preoperative clinical physical measurements and airway ultrasound parameters showed some discrimination for difficult airway. Conventional airway assessment measures and airway ultrasound parameters provide complementary sources of information. When combined with machine learning methods, they can more fully capture the joint value of multiple variables and offer an analytical approach to preoperative difficult airway prediction. This study provides a new reference for early identification and individualized airway management and lays a foundation for subsequent larger, multicenter clinical validation studies.

List of abbreviations

Abbreviation	Full term
Acc	Accuracy
AIC	Akaike information criterion
AP	Average precision

ASA	American Society of Anesthesiologists
AUC	Area under the curve
BAcc	Balanced accuracy
BIC	Bayesian information criterion
BMI	Body mass index
CI	Confidence interval
ECG	Electrocardiography
EtCO2	End-tidal carbon dioxide
FN	False negatives
FP	False positives
GAM	Generalized additive model
HMDR	Hyomental distance ratio
LRT	Likelihood-ratio test
NBP	Noninvasive blood pressure
NPV	Negative predictive value
PPV	Positive predictive value
PR	Precision–recall
PR-AUC	Area under the precision–recall curve
ROC	Receiver operating characteristic
ROC-AUC	Area under the receiver operating characteristic curve
SED	Skin-to-epiglottis distance

Sen	Sensitivity
SMD	Sternomental distance
SMOTE	Synthetic minority oversampling technique
Spe	Specificity
SpO2	Peripheral oxygen saturation
TN	True negatives
TP	True positives
TT	Tongue thickness
XGBoost	Extreme gradient boosting

Declarations

- Ethics approval and consent to participate

  Approval was obtained from the Medical Ethics Committee of the Second Affiliated Hospital of Zhengzhou University (approval no. KY2026232).

  The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

  Written informed consent was obtained from all participants before enrollment.

- Consent for publication

  Written informed consent for publication of clinical data and images was obtained from all participants.

- Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

- Competing interests

The authors declare that they have no competing interests.

- Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

- Authors' contributions

Jinghao Sun contributed to methodology, investigation, data curation, and manuscript drafting and revision. Li Xia contributed to conceptualization and supervision. Tianhao Sun contributed to methodology, formal analysis, and manuscript drafting and revision. All authors read and approved the final manuscript.

- Acknowledgements

No applicable.

Additional files

File name: Additional file 1

File format: .docx

Title: Supplementary tables for difficult airway prediction

Description: Table S1 presents descriptive statistics for the seven predictors in the training set. Table S2 summarizes the types and counts of difficult airway events. Table S3 reports model

performance during hyperparameter tuning using internal five-fold cross-validation in the training set. Table S4 presents restricted cubic spline tests for nonlinearity in the training set.

Reference

1 Cook TM, Oglesby F, Kane AD, Armstrong RA, Kursumovic E, Soar J. Airway and respiratory complications during anaesthesia and associated with peri-operative cardiac arrest as reported to the 7th national audit project of the royal college of anaesthetists. Anaesthesia. 2024 Apr;79(4):368–79.

2 Cumberworth A, Lewith H, Sud A, Jefferson H, Athanassoglou V, Pandit JJ. Major complications of airway management: A prospective multicentre observational study. Anaesthesia. 2022 June;77(6):640–8.

3 Apfelbaum JL, Hagberg CA, Connis RT, Abdelmalak BB, Agarkar M, Dutton RP, et al. 2022 American Society of Anesthesiologists Practice Guidelines for Management of the Difficult Airway. Anesthesiology. 2022 Jan;136(1):31–81.

4 Ahmad I, El-Boghdadly K, Iliff H, Dua G, Higgs A, Huntington M, et al. Difficult airway society 2025 guidelines for management of unanticipated difficult tracheal intubation in adults. Br J Anaesth. 2026 Jan;136(1):283–307.

5 Hung K-C, Chuang M-H, Kang F-C, Chang Y-J, Lin C-M, Yu C-H, et al. Prevalence and risk factors of difficult mask ventilation: A systematic review and meta-analysis. J Clin Anesth. 2023 Nov;90:111197.

6 Tang Z, Chen Q, Huang W, Wang J, Zou X, Xiao Y, et al. Difficult airway management in 25 hospitals across China: A multicenter cross-sectional study. J Clin Anesth. 2025 Mar;102:111766.

7 Soni VM, Pappu A, Zarabi S, Khalil C, You-Ten KE, Siddiqui N, et al. Point-of-care ultrasound of the upper airway in difficult airway management: A systematic review and meta-analysis*. Anaesthesia. 2025 Dec;80(12):1556–67.

8 Sajayan A, Nair A, McNarry AF, Mir F, Ahmad I, El-Boghdadly K. Analysis of a national difficult airway database. Anaesthesia. 2022 Oct;77(10):1081–8.

9 Giordano G, Alessandri F, Zulian A, Bilotta F, Pugliese F. Pre-operative ultrasound prediction of difficult airway management in adult patients: A systematic review of clinical evidence. Eur J Anaesthesiol. 2023 May;40(5):313–25.

10 Bhargava V, Rockwell NA, Tawfik D, Haileselassie B, Petrisor C, Su E. Prediction of difficult laryngoscopy using ultrasound: A systematic review and meta-analysis. Crit Care Med. 2023 Jan;51(1):117–26.

11 Carsetti A, Sorbello M, Adrario E, Donati A, Falcetta S. Airway ultrasound as predictor of difficult direct laryngoscopy: A systematic review and meta-analysis. Anesth Analg. 2022 Apr;134(4):740–50.

12 Sorbello M, La Via L, Paternò DS, Tutino S, Lo Giudice EC, Lentini M, et al. Artificial intelligence in airway management: A narrative review. Br J Anaesth. 2026 May;136(5):1606–15.

13 Demir Senoglu G, Kara Timucin E, Kırısoglu S. Evaluation of anthropometric and ultrasonographic measurements with different machine learning methods in predicting difficult intubation: A prospective observational study. BMC Anesthesiol. 2025 Oct;25(1):496.

14 Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. BMJ. 2021 Oct;n2281.

15 Altınsoy KE, Bayhan BU. Ultrasound-measured skin-to-epiglottis distance as a predictor of difficult intubation in obese patients: A prospective observational study. J Clin Med. 2025 Mar;14(6):2092.

16 Rastogi A, Singh AK, Srivastava D, Kannaujia AK, Singh TK, Mishra P. Evaluation of ultrasound measured tongue thickness, tongue thickness–thyromental distance ratio, and skin-to-epiglottis distance in predicting unanticipated difficult laryngoscopy. J Anaesthesiol Clin Pharmacol. 2025 Jan;41(1):151–7.

17 Furia JS, Nadkarni MM. Ultrasound-based assessment of tongue thickness for prediction of difficult laryngoscopy and intubation. J Anaesthesiol Clin Pharmacol. 2024 Apr;40(2):235–41.

18 Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: A step-by-step guide. BMJ. 2024 Sept;386:e078276.

19 Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation

of clinical prediction models (part 1): From development to external validation. BMJ. 2024 Jan;384:e074819.

20 Lopez-Ayala P, Riley RD, Collins GS, Zimmermann T. Dealing with continuous variables and modelling non-linear associations in healthcare data: Practical guide. BMJ. 2025 July;390:e082440.

21 Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): How to undertake an external validation study. BMJ. 2024 Jan;384:e074820.

22 Van Calster B, Collins GS, Vickers AJ, Wynants L, Kerr KF, Barreñada L, et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance. Lancet Digit Health. 2025 Dec;7(12):100916.

23 Clifton L, Powell J, Clifton DA, Sheikh A. When to and when not to use machine learning in risk prediction models. Lancet Digit Health. 2026 Feb;8(2):100954.

24 Volovici V, Syn NL, Ercole A, Zhao JJ, Liu N. Steps to avoid overuse and misuse of machine learning in clinical research. Nat Med. 2022 Oct;28(10):1996–9.

25 Riley RD, Collins GS, Kirton L, Snell KI, Ensor J, Whittle R, et al. Uncertainty of risk estimates from clinical prediction models: Rationale, challenges, and approaches. BMJ. 2025 Feb;e080749.

595 26 Thomassen D, Hackmann T, Goeman J, Steyerberg E, Le Cessie S. Effective sample
596 size for individual risk predictions: Quantifying uncertainty in machine learning
597 models. Lancet Digit Health. 2025 Nov;7(11):100911.

598 27 Riley RD, Ensor J, Snell KIE, Archer L, Whittle R, Dhiman P, et al. Importance of
599 sample size on the quality and utility of AI-based prediction models for healthcare.
600 Lancet Digit Health. 2025 June;7(6):100857.

601 28 Cloyd BH, Leis AM, Healy DW, Schechtman SA, Teig M. A retrospective
602 observational study of airway management features resulting in difficult airway
603 letters at a single center. J Clin Anesth. 2022 Dec;83:110984.

604 29 Huitink JM. Airway triage: A novel application-based method for airway assessment
605 and risk stratification. Br J Anaesth. 2024 May;132(5):987–90.

606 30 Giddings R, Joseph A, Callender T, Janes SM, Van Der Schaar M, Sheringham J, et al.
607 Factors influencing clinician and patient interaction with machine learning-based
608 risk prediction models: A systematic review. Lancet Digit Health. 2024
609 Feb;6(2):e131–44.

610

Supplementary Files

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

- Additionalfile1.docx