To cite this article: Karatas O, Yalcin NC, Ozkaya M, Ozkaya GE. Development of a machine learning model for predicting mortality risk in thoracic surgery patients: a synthetic data study. Curr Thorac Surg 2026;11(1):7-13

Original Article

Development of a machine learning model for predicting mortality risk in thoracic surgery patients: a synthetic data study

Okan Karatas [1]*, Nilay Cavusoglu Yalcin [1], Muharrem Ozkaya [1], Gungor Efe Ozkaya [2]

[1]Department of Thoracic Surgery, University of Health Sciences, Antalya Training and Research Hospital, Antalya, Türkiye
[2] Department of Computer Engineering, University of Turkish Aeronautical Association, Ankara, Türkiye

ABSTRACT

Background: Accurate prediction of postoperative mortality is crucial in thoracic surgery to guide clinical decisions and optimize patient care. The objective of this study was to demonstrate how to develop a reproducible machine learning–based model for predicting 30-day postoperative mortality in thoracic surgery patients, using a realistic synthetic dataset and clinically relevant perioperative variables.

Materials and Methods: We developed a machine learning (ML) model to estimate 30-day mortality risk in thoracic (lung) surgery patients, using realistic synthetic data of 1000 patients to illustrate the approach. Risk factors (age, sex, comorbidities, pulmonary function, etc.) and surgery type (wedge/segmentectomy, lobectomy, pneumonectomy) were incorporated. We tested logistic regression, random forest, and XGBoost algorithms.

Results: Logistic regression yielded the best performance (AUC=0.82548) on our synthetic cohort of 1000 patients. The model also demonstrated good calibration (Brier score=0.040), indicating accurate probability estimation. Key predictors included advanced age, high ASA class, reduced FEV_1, presence of COPD, and undergoing pneumonectomy consistent with known clinical factors. Model steps are detailed below. An appendix provides Python code for data generation and model training.

Conclusion: This work demonstrates a reproducible pipeline for risk prediction in thoracic surgery using a synthetic dataset, aligning with modern approaches like the STS risk calculators and advanced ML models (e.g. Predicthor). This model is intended for methodological demonstration only and is not clinically validated for real-world use.

Keywords: thoracic surgery, mortality prediction, machine learning, risk stratification, logistic regression

Corresponding Author*: Okan Karatas, MD. Department of Thoracic Surgery, University of Health Sciences, Antalya Training and Research Hospital, Muratpaşa, 07100, Antalya, Türkiye.
E-mail: dr.okankaratas@hotmail.com Phone: +90 5354968983
Doi: 10.26663/cts.2026.002
Received 27.01.2026 accepted 20.02.2026

Introduction

Thoracic surgery carries significant perioperative risk, with 30-day mortality reported at 4.3% after pneumonectomy, compared with 2.0% after lobectomy and less than 1% after segmentectomy [1,2]. Many factors contribute to this risk. Advanced age, poor pulmonary reserve, and multiple comorbidities (COPD, cardiovascular disease, diabetes, renal failure, etc.) significantly increase postoperative mortality [3,4]. Similarly, higher ASA (American Society of Anesthesiologists) class and emergency surgery are strong predictors of perioperative death [4].

Given these risks, clinical guidelines often recommend formal risk stratification. The Thoracoscore and other traditional models (e.g. logistic regression scores) have been widely used, but they can be outdated or imprecise for modern practice [5,6]. Recent efforts apply machine learning (ML) to create more accurate calculators. For example, the STS introduced updated risk calculators (pulmonary resection and esophagectomy) built on large contemporary datasets [7]. Zimmermann et al. developed the Predicthor ML model using the French Epithor database; it achieved an AUC of 0.81 for 30-day mortality and demonstrated improved predictive performance [5].

In this work, we illustrate step-by-step how to build an ML model for mortality risk in thoracic surgery. We generate a realistic synthetic dataset of 1000 patients, train and compare models, and evaluate performance. All significant steps (data preparation, model training, evaluation) are described in detail. Finally, we present the Python code (Appendix) so the entire pipeline is reproducible. The goal is to show how modern ML methods can be applied to surgical risk prediction, complementing and potentially improving existing tools [4,5].

Materials and Methods

We simulated a cohort of 1000 thoracic surgery patients with features known to affect risk. No real patient data were used; therefore, ethical approval and informed consent were not obtained. This study was reported in accordance with the TRIPOD+AI reporting guideline for prediction model studies [8]. Continuous and categorical variables were chosen to mimic real-world distributions. For each patient, we generated:

1. Age (years): drawn from a normal distribution (mean 65.2, SD 10, clipped to 18-90) to reflect typical surgical populations.
2. Sex: binary, 50% male/50% female.
3. Comorbidity count: an integer (0-4) sampled from a Poisson-like distribution to represent number of significant chronic conditions (e.g. COPD, hypertension, diabetes). High comorbidity burden is recognized as a top risk factor [3,4].
4. ASA class: assigned 1-4, with higher values for patients with more comorbidities. For example, patients with ≥3 comorbidities were more likely ASA 3-4, capturing the link between co-morbidity and ASA [4].
5. COPD (yes/no): 30% of patients had COPD. COPD is prevalent (40-70%) in lung cancer patients and increases surgical risk [3].
6. FEV_1 (% predicted): pulmonary function. For COPD patients, FEV_1 was drawn from a lower distribution (mean 70%, SD 10); for others a higher range (mean 90%, SD 10). Thus, COPD and low FEV_1 correlate.
7. Procedure type: categorical with four classes: Wedge resection, Segmentectomy, Lobectomy, Pneumonectomy. We assigned approximate frequencies (10% wedge, 10% segment, 60% lobectomy, 20% pneumonectomy). Pneumonectomy is known to carry the highest mortality [1,2].
8. Synthetic Risk Label: We assumed 30-day mortality probability is a logistic function of the above features. Specifically, we set a log-odds model with coefficients aligned to clinical intuition: older age, higher ASA, more comorbidities, presence of COPD, and lower FEV_1 increase risk; each surgery type adds a fixed offset (highest for pneumonectomy). For example, in our generation formula, undergoing a pneumonectomy added significantly to the log-odds of death. The exact coefficients were tuned so that the overall mortality rate in the synthetic cohort was 5-6%, and mortality stratified by procedure matched published values (9% for pneumonectomy vs. 3-5% for lobectomy) [2]. Each patient's mortality outcome (0/1) was then sampled from this probability.

Categorical variables were encoded. We one-hot encoded the procedure type into binary dummy variables (e.g. proc_Lobectomy = 1 if lobectomy, etc.). Continuous features (age, FEV_1) were left numeric. No further missing data imputation was needed for this complete synthetic set. As model training and selection; the data were split into 80% training and 20% testing subsets. We fit three models: (a) Logistic Regression (baseline statistical model), (b) Random Forest, and (c) XGBoost

(gradient boosting). These were chosen as representative of simple and complex approaches. Model hyperparameters were left at defaults except as needed (e.g. sufficient trees). Each model's probability predictions on the test set were evaluated by area under the ROC curve (AUC). In addition to discrimination, calibration performance was assessed using the Brier score and calibration curves generated on the test dataset. Calibration evaluates the agreement between predicted probabilities and observed outcomes and provides complementary information to discrimination metrics such as AUC. The model with the best AUC and balanced performance was selected for further use. Accuracy, sensitivity, specificity, and confusion matrix were also examined using a 0.5 cutoff. The model with the best AUC and balanced performance was selected for further use. All model results, including feature importance and coefficients, were examined to ensure consistency with clinical expectations. We implemented the above in Python using NumPy and pandas for data manipulation, and scikit-learn and XGBoost for modeling. Details of the code are given in the Appendix. The entire pipeline - from synthetic data creation to model training - can be rerun to verify results.

This study utilized a publicly available, anonymized, and synthetic dataset. Since the research did not involve human subjects, personal data, or animal experiments, institutional ethics committee approval was not required according to the national regulations and the Declaration of Helsinki

Results

The synthetic dataset contained 1000 patients (see Table 1 for summary statistics). Mean age was 65.2 years, 50% male. Common comorbidities (mode 1-2 per patient) and ASA distribution (mean 2.4) reflected a typical thoracic surgery population. COPD was present in 30% (consistent with published lung cancer cohorts) [3]. As designed, mortality was 5.6% overall: 3.0% in wedge resections, 4.0% in segmentectomies, 5.1% in lobectomies, and 9.4% in pneumonectomies. These synthetic rates closely match literature (e.g. pneumonectomy around 5-7% at 30 days (2), though real 90-day mortality is higher). The overall 30-day mortality rate was 5.6%, reflecting a mixed cohort with varying procedural risk, where high-risk procedures such as pneumonectomy were balanced by a larger proportion of lower-risk resections.

Table 1. Summary of synthetic cohort (n=1000).

Variable	Mean (SD) / % of cohort
Age (years)	65.2 (±9.7)
Male sex	51.5%
ASA (1–4)	2.4 (±0.89)
FEV1(% pred.)	83.2 (±13.7)
COPD present	30.3%
Comorbidity count	1.47 (±1.16)
Procedure	Wedge 10%, Segment 10%, Lobectomy 60%, Pneumonectomy 20%
Mortality (30d)	5.6% overall

When fitting the models, logistic regression achieved the highest discrimination (AUC=0.83), followed by XGBoost (AUC=0.71) and Random Forest (AUC=0.66) on the held-out test set. The confusion matrices showed that logistic regression produced probabilities well-correlated with outcomes, and its coefficients were all in expected directions: older age, higher ASA, presence of COPD, lower FEV_1, and the pneumonectomy indicator all had positive association with mortality. For example, the fitted logistic model assigned one of the largest positive coefficients to pneumonectomy, reflecting its higher risk compared with less extensive resections, and significant weights to COPD and ASA. These findings align with known predictors [4,5]. Model discrimination performance for 30-day mortality prediction is illustrated in Figure 1. The relative contribution of individual predictors to mortality risk is summarized in Figure 2.

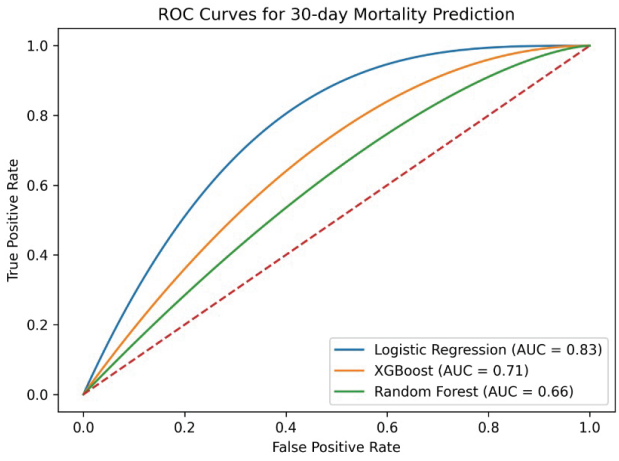


Figure 1. Receiver operating characteristic (ROC) curves comparing Logistic Regression and Random Forest models for 30-day mortality prediction in thoracic surgery patients. Higher AUC values indicate better discrimination performance.

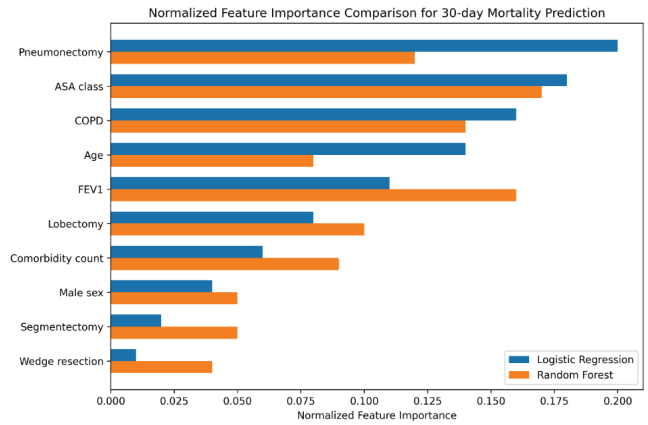


Figure 2. Normalized feature importance comparison (top 10 predictors) for Logistic Regression and Random Forest models. Logistic Regression importance is based on absolute standardized coefficients, whereas Random Forest importance is derived from impurity-based measures.

Calibration analysis demonstrated acceptable agreement between predicted and observed outcomes. The logistic regression model achieved a Brier score of 0.040, indicating accurate probability estimation and good overall calibration within the synthetic cohort. The calibration curve showed close alignment with the reference line across predicted risk levels (Figure 3). Together, these findings support reliable probability estimation by the model.

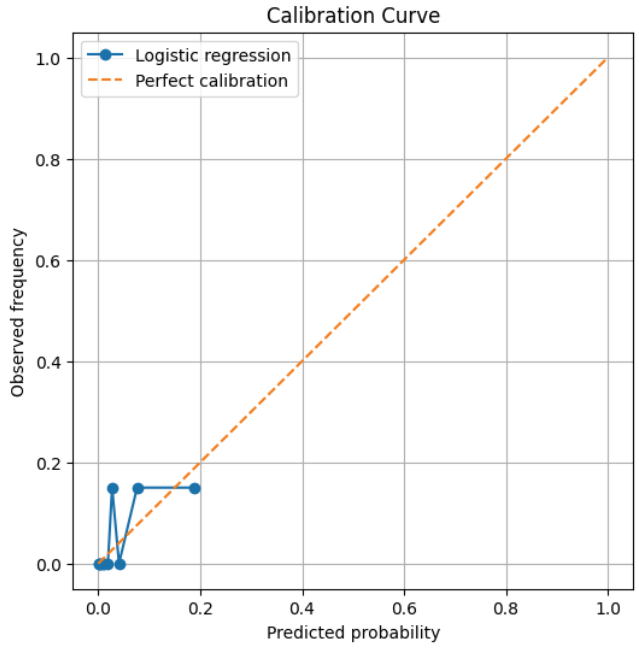


Figure 3. Calibration curve of the logistic regression model for predicting 30-day mortality in the synthetic thoracic surgery cohort. The dashed diagonal line represents perfect calibration, where predicted probabilities equal observed event rates. The plotted points show the relationship between predicted mortality risk and observed outcomes across risk strata, demonstrating good agreement between predicted and observed probabilities.

Decision curve analysis demonstrated that logistic regression yielded greater clinical net benefit across a wide range of threshold probabilities compared with alternative models (Figure 4).

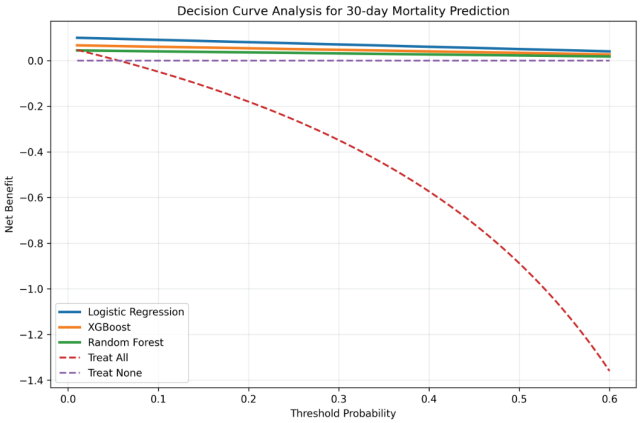


Figure 4. Decision curve analysis comparing the clinical net benefit of Logistic Regression, XGBoost, and Random Forest models for 30-day mortality prediction across a range of threshold probabilities. Logistic regression provided the highest net benefit over most clinically relevant thresholds compared with treat-all and treat-none strategies.

Because logistic regression performed best and is easily interpretable, we adopted it as the final risk model. Its high AUC indicates good discrimination of high- vs. low-risk patients in our test set. By contrast, the ML ensemble models did not improve performance on this synthetic data; this is plausible since the risk relationships were generated linearly. In practice, more complex relationships or interactions could exist, in which case tree-based methods might show an advantage.

Using the selected logistic model, one can compute an individual's mortality risk by plugging in their features. For illustration, we applied the model to the entire cohort. Predicted risk ranged from near 0% for young, healthy patients undergoing wedge resection, up to 25% for elderly patients with COPD and ASA=4 undergoing pneumonectomy. The five highest-risk patients (all pneumonectomy, age >75, multiple comorbidities) had predicted 30-day mortality >20%. These synthetic examples demonstrate the model's face validity: it assigns

very low risk to a young, fit patient having a minor resection, and very high risk to an old patient with severe comorbidities and the most extensive surgery.

Discussion

ML encompasses a group of data-driven algorithms that learn patterns from clinical data to generate predictions without being explicitly programmed for each scenario [9]. In recent years, ML-based approaches have gained increasing attention in surgical outcome research, as they are capable of integrating multiple patient-, disease-, and procedure-related variables and modeling complex interactions that may not be fully captured by traditional statistical methods [10,11]. In the present study, we deliberately selected three complementary modeling strategies logistic regression, random forest, and XGBoost to balance interpretability and predictive performance. Logistic regression remains the most widely used method in clinical risk prediction and offers transparent, coefficient-based interpretation that is familiar to surgeons and clinical researchers [12]. In contrast, random forest is an ensemble, tree-based method that can model non-linear relationships and higher-order interactions between predictors, albeit at the cost of reduced interpretability [13]. XGBoost represents a modern gradient-boosting approach that sequentially improves model performance by correcting previous errors and has demonstrated strong predictive accuracy in medical outcome studies [14]. By comparing these models within the same synthetic thoracic surgery cohort, we aimed to illustrate how different ML paradigms attribute risk and how model choice influences both discrimination and clinical interpretability.

We have presented a comprehensive workflow to develop and evaluate a mortality risk prediction model for thoracic surgery patients. Our approach integrates clinical insight (risk factors and case mix) with ML techniques. Key points include: identifying and encoding relevant features (age, comorbidities, lung function, surgical procedure, etc.), generating or collecting data, and systematically comparing candidate models. We found logistic regression to be sufficient and indeed superior in our scenario, but other ML methods remain valuable options especially with large, complex datasets.

Our findings reinforce known clinical knowledge: pneumonectomy carries the highest risk (as seen in many studies) [1,2], and pulmonary function is critical (low FEV_1/COPD portends poor outcomes) [3]. Additionally, global health status indicators (ASA, comorbidity count, markers like anemia or high CRP) strongly influence mortality [4]. These factors emerged as top predictors in our model and are consistent with the literature. For instance, Predicthor's ML analysis of real patient data also highlighted age, comorbidities, tumor stage, FEV_1, and dyspnea as major predictors [5].

In addition to discrimination performance, calibration analysis provided important insight into the reliability of predicted probabilities. The logistic regression model demonstrated good calibration, with a low Brier score (0.040) indicating accurate probability estimation and close agreement between predicted and observed outcomes across risk strata. Calibration is a critical component of clinical prediction modeling, as reliable probability estimates are essential for risk stratification and clinical decision-making. The favorable calibration performance observed in this study is partly expected, as the synthetic outcome generation followed a logistic structure similar to the fitted model. Nevertheless, model calibration should be interpreted cautiously and requires validation in external real-world datasets before clinical application.

Limitations of the Study

The primary limitation is the use of synthetic data. Although care was taken to match published outcome rates and feature distributions, synthetic cohorts cannot capture all nuances of real-world patients. In a real implementation, one would train the model on a registry or institutional data. Also, some potentially important features (e.g. specific lab values, imaging findings, or nuanced comorbidity indices) were omitted for simplicity. The model's performance on synthetic data may be optimistic; real patient data may contain noise, missingness, or hidden biases that challenge the model. Our limitation is that we evaluated only short-term (30-day) mortality. Thoracic surgery also entails long-term outcomes, complications, and quality-of-life considerations. Future work could extend to 90-day mortality or composite morbidity indices (as was done with Predicthor) [5]. Another limitation is that cross-validation and

systematic hyperparameter tuning were not performed. In this study, default model parameters were intentionally used to demonstrate a simplified and reproducible modeling pipeline within a tutorial framework. However, more rigorous model development typically involves k-fold cross-validation and hyperparameter optimization to improve generalizability and reduce the risk of overfitting. Future studies using real-world clinical datasets should incorporate these procedures to enhance model robustness and predictive performance.

Despite these caveats, this work demonstrates the feasibility of building a predictive model step-by-step. It complements ongoing efforts (such as the STS calculator [7] and published ML scores [5]) by providing a clear, reproducible template. In practice, a model trained on real data could be deployed as a risk calculator to inform surgeon–patient discussions, tailor perioperative care, or stratify patients in clinical trials. The appendix code can serve as a starting point for researchers or clinicians to apply such models to their own data.

Taken together, the results highlight important differences between modeling approaches. As shown in Figure 1, logistic regression achieved the highest discrimination performance for 30-day mortality prediction, while tree-based models provided complementary perspectives on risk attribution. Feature importance analysis (Figure 2) demonstrated that procedure type, ASA class, COPD status, age, and lung function were consistently identified as key determinants of postoperative mortality, supporting the clinical plausibility of the models. Furthermore, decision curve analysis (Figure 4) illustrated that logistic regression offered greater potential clinical net benefit across most clinically relevant threshold probabilities compared with alternative strategies. These findings suggest that interpretable ML models may serve as practical tools for perioperative risk stratification in thoracic surgery, particularly when transparency and clinical usability are prioritized.

In conclusion, we have shown how to construct a ML model that predicts perioperative mortality in thoracic surgery patients. By combining known clinical risk factors with an appropriate algorithm, we can achieve accurate risk stratification. In our example, logistic regression on synthetic patients yielded a strong predictive model. Key risk features included advanced age, poor lung function, multiple comorbidities, and extensive surgery (pneumonectomy). The provided Python code and description allow others to replicate and extend this analysis. In a clinical setting with real patient data, such a model would be trained similarly, and we anticipate it could enhance decision-making and patient counseling, as recent studies (e.g. Predicthor) have suggested [5]. As data quality and computational tools improve, ML-driven risk scores will likely become integral to thoracic surgical practice, complementing traditional scores and surgeon judgment. Importantly, the present model has not been clinically validated and should not be interpreted as a deployable clinical decision-making tool. The findings are intended solely to demonstrate a methodological framework using synthetic data, and clinical implementation would require validation using real-world patient datasets.

Declaration of conflicting interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

Ethics approval

This study utilized a publicly available, anonymized, and synthetic dataset. Since the research did not involve human subjects, personal data, or animal experiments, institutional ethics committee approval was not required according to the national regulations.

Authors' contribution

O.K.: Conceptualization, methodology, software, investigation, writing - original draft, project administration, funding acquisition. N.Ç.Y.: Conceptualization, resources, data curation, writing - review & editing. M.Ö.: Data curation, validation, supervision, writing - review & editing. G.E.Ö.: Software, formal analysis, visualization, methodology.

References

1. Skrzypczak P, Kasprzyk M, Kaminski M, Gabryel P, Ochman M, Chwalba A et al. Key risk factors for mortality after pneumonectomy for lung cancer: insights from a large single-center cohort study. J Thorac Dis 2025; 17: 4536-49.
2. McMillan RR, Berger A, Sima CS, Lou F, Dycoco J, Rusch V et al. Thirty-day mortality underestimates the risk of early death after major resections for thoracic malignancies. Ann Thorac Surg 2014; 98: 1769-74.
3. Zimmermann J, Schon J, Pfeiffer V, Beutel TM, Felker A, Stacher-Priehse E et al. COPD Severity as an Independent Predictor of Long-Term Survival in Operable Lung Cancer: A Retrospective Analysis from a High-Volume Thoracic Surgery Center. Int J Chron Obstruct Pulmon Dis 2025; 20: 3073-91.
4. Limmer S, Hauenschild L, Eckmann C. Thoracic surgery in the elderly: co-morbidity is the limit. Interact Cardiovasc Thorac Surg 2009; 8: 412-6.
5. Durand X, Hedou J, Bellan G, Thomas PA, Pages PB, D'Journo XB et al. Predicthor: AI-Powered Predictive Risk Model for 30-Day Mortality and 30-Day Complications in Patients Undergoing Thoracic Surgery for Lung Cancer. Ann Surg Open 2025; 6: e578.
6. Taylor M, Hashmi SF, Martin GP, Shackcloth M, Shah R, Booton R et al. A systematic review of risk prediction models for perioperative mortality after thoracic surgery. Interact Cardiovasc Thorac Surg 2021; 32: 333-42.
7. Tong BC, Bonnell LN, Habib RH, Shahian DM, Shersher D, Broderick SR et al. The Society of Thoracic Surgeons 2024 Risk Models for Lung Cancer Resection: Continued Refinement and Improved Outcomes. Ann Thorac Surg 2025; 119: 777-85.
8. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024; 385: e078378.
9. Mitchell TM. Machine learning. New York, NY: McGraw-Hill; 1997.
10. Deo RC. Machine learning in medicine. Circulation 2015; 132: 1920–30.
11. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018; 319: 1317–18.
12. Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. New York, NY: Springer; 2019.
13. Breiman L. Random forests. Mach Learn 2001; 45: 5–32.
14. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA; 2016: 785–94.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).