Original Study

ANNALS OF SURGERY OPEN

OPEN

Predicthor

AI-Powered Predictive Risk Model for 30-Day Mortality and 30-Day Complications in Patients Undergoing Thoracic Surgery for Lung Cancer

Xavier Durand, Msc,* Julien Hédou, Msc,*,† Grégoire Bellan, Msc,* Pascal-Alexandre Thomas, MD, PhD,‡ Pierre-Benoît Pages, PhD,§ Xavier-Benoît D'Journo, PhD,‡ Laurent Brouchet, PhD,‖ Caroline Rivera, MD,¶ Pierre-Emmanuel Falcoz, PhD,# André Gillibert, MD, PhD,** Jean-Marc Baste, MD, PhD**

Objective: To assess the predictive performance of Predicthor, an artificial intelligence model, for 30-day mortality and complications following major pulmonary resections.

Background: The significance of predicting postoperative complications in thoracic surgery lies in the impact on patient outcomes and the efficient allocation of healthcare resources. The longstanding use of the Thoracoscore for over 15 years in hospital settings emphasizes the opportune moment for an update, leveraging new artificial intelligence methodologies to enhance predictive precision and relevance.

Methods: The EPITHOR French population-based database linked to the National Institute of Statistics and Economic Studies database has been queried from January 1, 2016, through December 31, 2022, on 6 selected hospital centers (Rouen, Dijon and Toulouse CHUs, Strasbourg CHRU, Centre Hospitalier Général de Bayonne, and Assitance Publique des Hopitaux de Marseille) with curated data collection. A total of 6508 patients who have undergone primary lung cancer surgery via lobectomy or bilobectomy, aged over 18 years, and with anAmerican Society of Anesthesiologists (ASA) physical status classification system score under 4, were selected. In a retrospective analysis using a 3-dataset scheme (training cohort, internal and external validation on 118 other centers), we assessed the predictive performance of Predicthor for 30-day complications and mortality following major pulmonary resections.

Results: Postoperative complications occurred in 17.6% of patients, with 4.6% experiencing complications of Clavien–Dindo grade III or higher. Overall mortality was 0.6%. Predicthor excelled in predicting 30-day mortality with an area under the curve of 0.81 (95% CI = 0.79–0.83; $P < 1E{-}16$), surpassing the Thoracoscore at 0.72 (95% CI = 0.70–0.75; $P < 1E{-}16$). Predicthor identified 9 key variables, including age, comorbidity scores, tumor characteristics, forced expiratory volume (FEV1), and dyspnea. They were utilized for predicting Comprehensive Complication Index (Pearson-r: 0.23; 95% CI = 0.22–0.24; $P < 1E{-}16$) and complications with Clavien–Dindo ≥III (area under the curve: 0.68; 95% CI = 0.68–0.69; $P < 1E{-}16$).

Conclusions: Predicthor's predictive performance for 30-day mortality and complications highlighted its potential as a valuable tool in clinical decision-making. The study's methodology and comprehensive dataset contribute to its relevance in using machine learning on large available databases for shaping thoracic surgery practices and patient management.

Keywords: artificial intelligence, bilobectomy, Clavien–Dindo, complications, data, data curation, epithor, lobectomy, lung, machine learning, mortality, predicthor, surgery, thoracic surgery, thoracoscore, validation

INTRODUCTION

Surgery remains the cornerstone of treatment for various conditions, including early-stage cancers. However, postoperative complications persist as a significant concern, affecting up to 60% of surgical patients and resulting in increased morbidity, mortality, and healthcare costs.[1,2] Despite advancements in surgical techniques, the incidence of postoperative complications has not proportionally decreased.[3] Consequently, there is an emergent need for accurate predictive scores to preemptively stratify and identify patients' risk, enabling interventions to mitigate complications.

*From the *SurgeCare, SAS, Department of Data Science, Paris, France; †Sorbonne Université, Inserm, UMRS_938, Centre de Recherche Saint-Antoine (CRSA), Paris, France; ‡Department of Thoracic Surgery and Diseases of the Esophagus, Aix-Marseille University, Marseille, France; §Department of Thoracic Surgery, CHU Dijon, France; ‖Department of Thoracic Surgery, CHU Toulouse, Toulouse, France; ¶Department of Thoracic Surgery, Centre Hospitalier de la Côte Basque, Bayonne, France; #Department of Thoracic Surgery, CHU Strasbourg, Strasbourg, France; **Department of Cardio-Thoracic Surgery, CHU Rouen, Inserm, UNIVROUEN, France.*

Julien Hédou and Xavier Durand contributed equally to this study.

Disclosure: J.H. is a director; G.B. and X.D. are employed at SurgeCare. The other authors declare that there is nothing to disclose.

This study was conducted following approval from the French scientific committee Comité Scientifique en Chirurgie Thoracique et Cardio-Vasculaire. Given that this study is a retrospective analysis utilizing previously collected data, individual patient consent was not required. To protect patient privacy and confidentiality, all data were fully anonymized before analysis, in accordance with the ethical guidelines.

The data utilized in this study are sourced from the EPITHOR national database and can be made available upon reasonable and formal request, subject to approval by the corresponding author (jhedou@stanford.edu).

SDC *Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.annalsofsurgery.com).*

Reprints: Julien Hédou, Msc, Sorbonne Université, Inserm, UMRS_938, Centre de Recherche Saint-Antoine (CRSA), Paris, France. Email: jhedou@stanford.edu.

Annals of Surgery Open (2025) 2:e578

Received: 21 February 2025; Accepted 15 April 2025

Published online 27 May 2025

DOI: 10.1097/AS9.0000000000000578

In cardiothoracic surgery, predictive models like the Thoracoscore[4] have been acknowledged for their robust predictivity. Nonetheless, recent studies bring to light potential shortcomings in its performance.[4] Several attempts have been made to improve the Thoracoscore, but these efforts have not resulted in a significant performance enhancement for the adoption of a novel tool.[5–9] Historically, these models have focused primarily on mortality prediction, providing only a narrow view on morbidity. Predicting mortality alone may not be sufficiently informative for clinicians' decision-making, particularly when deliberating on surgical interventions or in optimizing preoperative health.

Moreover, the reliability and applicability of these models are compromised by the insufficient data quality control with which they have been trained. With the advent of minimally invasive surgery[10] and changes in the standard of care, there is a compelling case for a new, inclusive predictive tool that integrates enhanced data quality and updated algorithms to more comprehensively forecast both mortality and postoperative complications; to ultimately help surgeons to improve decision-making for frail patients.

Finally, scores and predictive models in the medical field, like those in cardiothoracic surgery, are not static entities; they require periodic updates to stay relevant and accurate. As medical practices evolve, incorporating new techniques, technologies, and treatment paradigms, the foundational data upon which these scores are built can become outdated. The quality and comprehensiveness of data in registries also improve over time. This continual refinement ensures the scores can remain a reliable tool for clinicians in decision-making and risk assessment.

The primary objective of this multicenter retrospective cohort study is to develop a robust and interpretable predictive score for postoperative complications and mortality within 30 days following major pulmonary resections (lobectomy and bilobectomy), named Predicthor. Utilizing comprehensive clinical data from the EPITHOR database,[11] we have imposed strict controls on data exhaustiveness and curation—elements overlooked in existing models. The predictive models were based either on in-hospital mortality or the Comprehensive Complication Index (CCI).

A distinguishing feature of Predicthor is its dual-layered validation methodology. The model was both internally and externally validated, with an independent run of results to assure robustness and reproducibility. Furthermore, Predicthor employs a unique and cutting-edge machine learning (ML) pipeline, aligning with the current effort of integrating artificial intelligence tools in the medical domain.[12,13] Not only does this pipeline optimize the model's predictive power, but it also derives an interpretable machine learning model. This allows for agnostic feature selection, enabling users to understand which variables significantly influence the predictions, thereby enhancing trust and transparency in the model's outputs. The model was compared with the Thoracoscore[4] and to a classic statistical approach where the final score is computed as a linear combination of variables selected by a domain expert.

Predicthor aims to set a standard for predictive models in cardiothoracic surgery by providing a more accurate predictive tool to enhance patient management. We introduce a novel model based on advanced machine learning methodologies trained with high-quality, rigorously curated data to offer a reliable and comprehensive predictive tool filling a long-standing gap in clinical practice.

METHODS

Study Participant Characteristics

The cohort used for this study is retrospective and has been selected from the national EPITHOR database[14] (Fig. 1A) and includes patients within the timeframe of January 1, 2016, to December 31, 2022. Inclusion criteria were patients who have undergone primary lung cancer surgery via lobectomy or bilobectomy (including completion lobectomy), aged over 18 years, and with an ASA score under 4. Six hospital centers [Centre Hospitalier Universitaire (CHU) de Rouen, CHU de Dijon, CHU de Toulouse, Centre Hospitalier Régional Universitaire de Strasbourg, Centre Hospitalier Général de Bayonne, and Assitance Publique des Hopitaux de Marseille (AP-HM)] were selected for the training set as they had data quality score over 90%, determined as the percentage of patients with a complete file. A data quality score over 90% means that over 90% of the records are completed and curated by a domain expert to verify that the fields are matching the patient medical record. The external validation set was performed on all other hospital centers in the EPITHOR database (118 centers).

Patients whose postresection pathological anatomy does not confirm the diagnosis are not excluded, as the prediction of postoperative complications is applicable before the definitive diagnosis is known. Pulmonary metastases from cancers of different origins are excluded if this diagnosis of metastasis was known before surgery. Multiple inclusions were permitted, meaning that the same patient could undergo 2 surgeries sequentially for 2 primary lung cancers and would then be counted as 2 cases. Thus, our analysis is based on the number of procedures performed rather than the number of individual patients.

Objectives

The primary objective of this research study is to employ an innovative machine-learning approach for the mortality prediction within 30 days following surgery. An integral aspect of this investigation included a comparative analysis of the model's predictive performance in relation to the previously trained Thoracoscore.[4]

Additionally, the study evaluated the model's potential in establishing a robust clinical predictive scoring system tailored to anticipate postoperative complications within 30 days after major pulmonary resections, including lobectomy and bilobectomy. While the 90-day mortality is increasingly used as a standard in thoracic surgery, a limitation in this study was the high rate of loss to follow-up beyond 30 days. Given this limitation, the 90-day mortality could not be assessed without introducing potential bias or inaccuracies. Therefore, we opted for the 30-day mortality threshold.

Comprehensive Complication Index

The CCI[16] is a scoring system used to quantify the overall burden of postoperative complications that a patient experiences after a surgical procedure. In the computation of the CCI score, we assigned specific weights to each grade of complication,[16] based on the severity determined by patients and physicians. The CCI score was calculated using the formula:

$$CCI = \frac{1}{2}\sqrt{\sum_{i}^{n} w_i},$$

where n is the number of complications and w_i the weight of the i-th complication. Notably, to maintain interpretability and ensure that the score did not exceed the maximum value of 100 (indicative of patient death), the overall score was clipped at 100.

Mortality

30-day mortality was defined as any death that occurred following surgery but before 30 days after the surgery. Deaths occurring subsequent to patient transfer from the surgical department to another medical-surgical department were also categorized as in-hospital deaths.

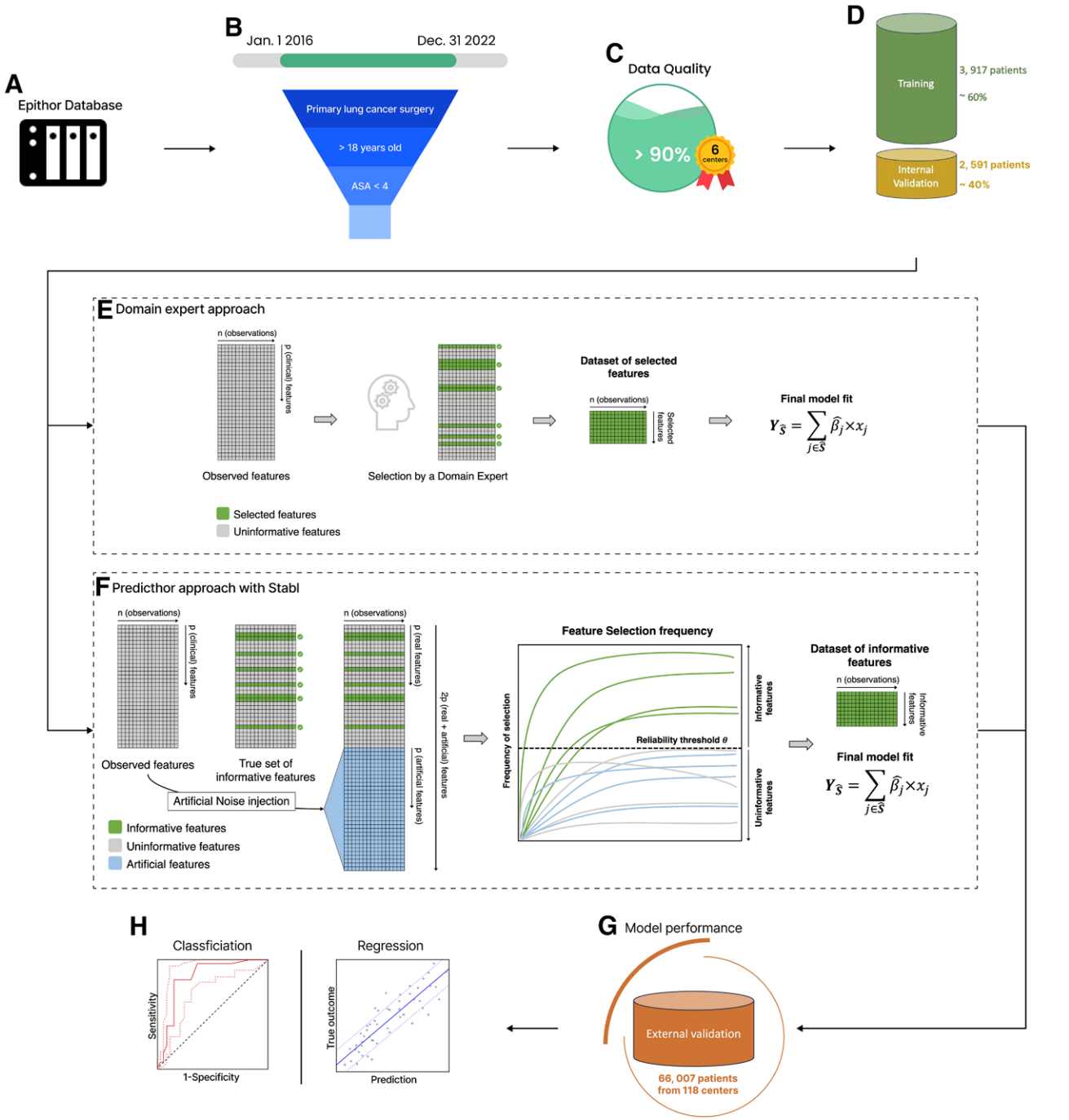

FIGURE 1. Overview of the training and model validation process. A–C, Dataset sourced from 6 centers with data quality over 90% within the Epithor database, adhering to predefined inclusion criteria. D, Dataset split into 2 subsets: the training set (60%; 3917 patients) and internal validation set (40%; 2591 patients). E, Sequential phases for model optimization, encompassing preprocessing with feature selection using expert knowledge, model training through weight generation, and model validation to identify the most suitable model. F, Predicthor training after Stabl feature selection: Informative features (green) and uninformative features (gray) are distinguished, with p artificial features (blue) added to create a dataset of size $n \times 2p$. The selection frequency $f_i(\lambda)$ of feature i is calculated over B subsample iterations, fitting SRM models with varying regularization parameters λ on each subsample, and plotted against 1/λ to create a stability path graph. Features with a maximum frequency above the threshold θ are selected for the final model. Stabl uses a reliability threshold θ, based on the FDP+,[15] to identify reliable features included in the final predictive model. A final regression is fitted on the selected set of features. G and H, Evaluation of prediction quality through external validation (66,007 patients) with area under ROC for classification and Pearson-R for regression.

Scale of Performance

In assessing the model quality in classification, a widely accepted performance scale was employed, wherein Receiver operating curve (ROC) area under the curve (AUC)[17] scores falling between 0.5 and 0.7 denoted fair performance, scores between 0.7 and 0.8 signified good performance, scores from 0.8 to 0.9 indicated very good performance, and scores from 0.9 to 1.0 were indicative of excellent performance. For regression analysis, the Pearson correlation coefficient was used to assess the performance using the scale where the absolute value of the coefficient under 0.2 corresponded to a poor performance, an absolute value between 0.2 and 0.5 depicted a good performance, an absolute value from 0.5 to 0.8 described a very good performance, absolute scores between 0.8 and 1.0 were considered as excellent performance.

Data Preprocessing

In the context of machine learning, data preprocessing plays a pivotal role in ensuring the effective training of models. While the selection of patients from the EPITHOR database adhered to stringent data quality criteria, some variables required specific preprocessing steps. Notably, continuous variables such as age, smoking status, and FEV1 underwent discretization to emphasize their significance in the analysis. Age was categorized into distinct groups: below 65 years, 65 to 75 years, 75 to 85 years, and above 85 years, reflecting the varying patterns of complications within each age bracket. Smoking status was dichotomous, indicating whether the patient was a smoker or not, while FEV1 was discretized into 4 distinct ranges: 0 to 50, signifying poor capacity; 50 to 75, indicative of fair capacity; 75 to 90, denoting good capacity; and above 90, representing excellent capacity. Similar transformations were performed on other variables and used during the feature selection process. They are not further detailed because they were not chosen.

Comorbidity Index and Comorbidity Score

Comorbidities were observed to exert a discernible influence on the occurrence of complications within the patient cohort. To comprehensively assess the impact of comorbidities on complications, two metrics were computed. The first, the comorbidity index, uses the same formula as in the CCI. This approach mirrors the weighted calculation applied to complications, with each comorbidity being assigned a specific grade based on its severity. The second metric, the comorbidity score, uses the numerical values associated with comorbidity grades. By performing the Euclidean norm of these grades, this metric quantifies the cumulative effect of comorbidities.

TNM Classification of Malignant Tumors

The TNM classification was standardized using the 3 variables: Tumor, Nodes, and Metastasis (TNM). To ensure data consistency, if 1 or more elements among the 3 were provided, the remaining variables were encoded as nonmissing with the letter 'x'. For example, if there was a preoperative tumor categorized as T3, N0, but metastasis was missing, it was recoded as Mx.

The preoperative TNM staging, which is either purely clinical or influenced by biopsy results, was recalculated based on T, N, and M variables, aligning as closely as possible with AJCCv7,[18] most commonly used in EPITHOR. The use of AJCCv8[19] was harmonized and adapted back to AJCCv7 to ensure consistency across the dataset (eg, T1c was recoded as T1b for compatibility with AJCCv7) using eTable 2, https://links.lww.com/AOSO/A500.

Whenever M1a or M1b was present, the stage was classified as IV. Cases classified as occult (due to Nx or Mx) were treated as missing data for TNM staging. Finally, if the detailed T, N, and M were unspecified but the overall TNM stage was available, it was used.

Missing Data Imputation

Addressing missing data is a pivotal aspect of data preprocessing,[20] as it plays a crucial role in upholding the dataset's completeness and integrity. Notably, for variables such as metastasis and prior thoracic surgery, where null values hold specific interpretative significance, they were deliberately retained to convey their intended meanings.

In the case of continuous variables, the chosen approach involved replacing missing values with the median of their respective variables. This imputation strategy is strategically employed to maintain the central tendency of the data, thus preventing any undue influence on subsequent analytical processes. Conversely, for categorical variables, the handling of missing data entailed their imputation with the majority class. This method ensured the preservation of the categorical structure and aligned imputed values with the most prevalent category within the dataset.

RESULTS

Study Participant Characteristics

The final cohort comprised 6508 patients meeting the inclusion criteria and was randomly divided into 2 datasets: a training cohort and internal validation (eTable 1, https://links.lww.com/AOSO/A500, Fig. 1A–D). The primary objective of this study was to harness machine learning techniques to construct a predictive model for postoperative complications, assessed by the CCI [16] and 30-day mortality. The CCI is a comprehensive continuous measure used to quantify the overall severity of postoperative complications. It ranges from 0 to 100, where higher scores indicate a severe impact of postoperative complications. In the dataset, it had an overall mean of 10.35 with a standard deviation of 17.11 and was composed of 3973 (61%) patients with no complications, and 2535 (39%) patients with complications.

The Clavien–Dindo classification,[21] highly correlated with the CCI, was used to assess the 30-day mortality, corresponding to 84 (1.3%) patients.

Modeling Strategy for the Prediction of Postoperative Complications and Mortality

To develop new models, 2 approaches were tested. The first one used classical statistical methods, where predictive variables were manually selected based on a combination of significant univariate statistical analysis and expert domain knowledge (Fig. 1E). The second one used Stabl,[15] a machine learning framework designed to identify a robust and concise set of biomarkers through multivariate predictive modeling during the learning phase. This method operates agnostically by injecting artificial features into the original dataset. The augmented dataset undergoes regularization techniques across multiple bootstrap procedures, producing a feature selection frequency graph. A reliability threshold is then established, above which real features are deemed informative. This threshold is optimized to minimize the inclusion of artificial features while maximizing the retention of informative ones (Fig. 1F). Both approaches were used to select variables, and a final regression model was used for prediction, allowing interpretability of the covariates selected in the models. Predicthor designates the model that is the result of Stabl best performance in cross-validation trained on the EPITHOR dataset.

According to the gold standard in machine learning model developments, and following the approach described notably in the Thoracoscore,[4] we defined a train-test strategy for picking the best model.[22] As a result, the model tuning was performed on the training (3917 patients) and internal validation datasets (2591 patients). An additional external validation (66,007 patients) was used to evaluate the final performance of each model. (Fig. 1G,H). Within this dataset, the CCI had a mean of 4.3 and a standard deviation of 11.9. The dataset included 10,979 patients (16.6%) who experienced complications, with 345 cases (0.37%) resulting in mortality.

Predicthor Outperforms Classic Statistics and Thoracoscore's Performance for the Prediction of Mortality

Prediction of mortality was performed using Predicthor, a custom model using a novel machine learning approach (Stabl),

and compared to classic statistics and the Thoracoscore.[4] On the external validation set, Predicthor outperformed the classic statistics and Thoracoscore models with a very good ROC AUC of 0.81 (95% CI = 0.79–0.83), 9 points above the Thoracoscore (95% CI = 0.70–0.75) and 12 points above the statistical approach (95% CI = 0.66–0.71) (Table 1 and Fig. 2). Additional performance metrics, which further support the superiority of Predicthor for mortality prediction, are presented in eTable 3, https://links.lww.com/AOSO/A500.

Beyond evaluating overall performance, we assessed how well each model's predictions aligned with real-world clinical expectations. To do this, we analyzed the distribution of predicted probabilities for both mortality (eFigure 1 https://links.lww.com/AOSO/A500) and survival (eFigure 2 https://links.lww.com/AOSO/A500) across different patient risk groups. The probability of mortality (eFigure 1 https://links.lww.com/AOSO/A500) followed a comparable trend in both the statistical model and Predicthor, gradually increasing before peaking at 3.2% and 12.5%, respectively. In contrast, Thoracoscore exhibited a lower maximum probability of 1.3%. These peak probabilities were observed in small patient subgroups, comprising 214 patients for the statistical model and only 16 for Predicthor, whereas Thoracoscore's peak probability was based on a larger subgroup of 8489 patients. Notably, with a similar patient count of approximately 250, Predicthor yielded a mortality probability twice as high as the statistical model, suggesting a more meaningful estimation. Conversely, the probability of survival demonstrated consistent trends across all 3 models (eFigure 2 https://links.lww.com/AOSO/A500).

Performance metrics revealed a similar AUC on the training set for both models [0.80 (95% CI = 0.75–0.85) and 0.79 (95% CI = 0.74–0.84), respectively, for the Predicthor and the statistical model]. This suggests a mild overfitting for the statistical model, suggesting a potential susceptibility to capturing minor fluctuations rather than robust, meaningful patterns. Conversely, Predicthor displayed consistent performance, indicating minimal to no overfitting.

To overcome the limited occurrence of mortality, this study broadened its scope to encompass the prediction of complication severity utilizing the CCI scale as an additional objective.

Extension of Machine Learning Models to the Capacity of Predicting Complications

The same approaches were carried out for the prediction of the continuous CCI score. While Predicthor outperformed the 2 other models with a Pearson correlation coefficient of 0.23, the statistical model and the Thoracoscore fell behind with poor performance of 0.08 and 0.06, respectively (Table 1). There was no overfitting of the models, with similar performance on the external cohort. A calibration analysis (eFigure 3, https://links.lww.com/AOSO/A500) was performed on the continuous task of predicting the CCI, and no significant calibration error was found.

The predictions obtained were also used to predict the presence of a severe complication. A complication was considered as severe when the Clavien–Dindo classification was not inferior to grade III. Predicthor outperformed both the classical approach and the Thoracoscore, achieving an ROC AUC of 0.68, compared to 0.60 for the classical approach and 0.57 for the Thoracoscore, which exhibited only fair predictive performance.

The Machine Learning Model Identified New Features Predictive of Mortality and Complications

Both methods had 3 features in common: the age category, the FEV1 category, and the preoperative tumor stage. The statistical model selected 4 specific features: the ECOG performance status, the ASA, either the patient was smoking and the type of intervention (lobectomy or bilobectomy); whereas the Predicthor approach selected 6 specific features: the M preoperative tumor stage, either there was a history of thoracic surgery, the dyspnea score, the comorbidity score, the comorbidity index, and the FEV1 as a continuous value (Table 2)

In summary, the statistical analysis identified 7 variables, while the machine learning approach extracted 9 variables. The same set of features was used to predict the CCI score in a regression task and mortality in a classification task.

Generalization to Other Intervention Types

Predicthor's model was initially trained on lobectomies using high-quality, hand-curated data from selected centers to ensure consistency. To assess its generalizability, we conducted a post-hoc analysis on 12,660 segmentectomy cases from the EPITHOR database, a procedure differing in surgical approach and complexity. Among these cases, 4.1% (n = 513) had high-grade complications, and 0.3% (n = 42) resulted in mortality. Predicthor maintained similar predictive performance to lobectomies, with AUROC = 0.79 (95% CI = 0.77–0.81) for mortality, AUROC = 0.66 (95% CI = 0.64–0.68) for high-grade complications, and a Pearson correlation of 0.21 (95% CI = 0.19–0.23) for CCI, as shown in eTable 4, https://links.lww.com/AOSO/A500.

To further strengthen our findings, we evaluated the 3 models on the full EPITHOR dataset (1985–2022; n = 302,484 patients). Within this cohort, 3.3% (n = 10,120) experienced high-grade complications (Clavien–Dindo ≥ III), and 0.6% (n = 1,783) resulted in mortality. For mortality prediction, Predicthor demonstrated superior performance with AUROC = 0.82 (95% CI = 0.81–0.83), outperforming Thoracoscore (AUROC = 0.79; 95% CI = 0.78–0.80) and statistical estimation (AUROC = 0.74; 95% CI = 0.73–0.75). For high-grade complications, Predicthor also outperformed the other models with AUROC = 0.73 (95% CI: 0.73–0.74), compared with Thoracoscore (AUROC = 0.67; 95% CI = 0.66–0.68) and statistical estimation (AUROC = 0.60; 95% CI = 0.59–0.60). Additionally, Predicthor showed the strongest correlation with CCI (Pearson r = 0.21; 95% CI = 0.20–0.22) (eTable 5,

TABLE 1.

Models for Analysis of Complications and Mortality on Test Set

	Statistical estimation (95% CI)	Machine learning estimation (95% CI)	Thoracoscore estimation (95% CI)
CCI, Pearson correlation	0.08 (0.08–0.09)	**0.23** (0.22–0.24)	0.06 (0.05–0.07)
Pearson *P* value	*P* < 1E–16	*P* < 1E–16	*P* < 1E–16
Clavien Dindo ≥ III, AUC	0.57 (0.56–0.58)	**0.68** (0.68–0.69)	0.60 (0.59–0.61)
Mann–Whitney *P* value	*P* < 1E–16	*P* < 1E–16	*P* < 1E–16
Death, AUC	0.69 (0.66–0.71)	**0.81** (0.79–0.83)	0.72 (0.70–0.75)
Mann–Whitney *P* value	*P* < 1E–16	*P* < 1E–16	*P* < 1E–16

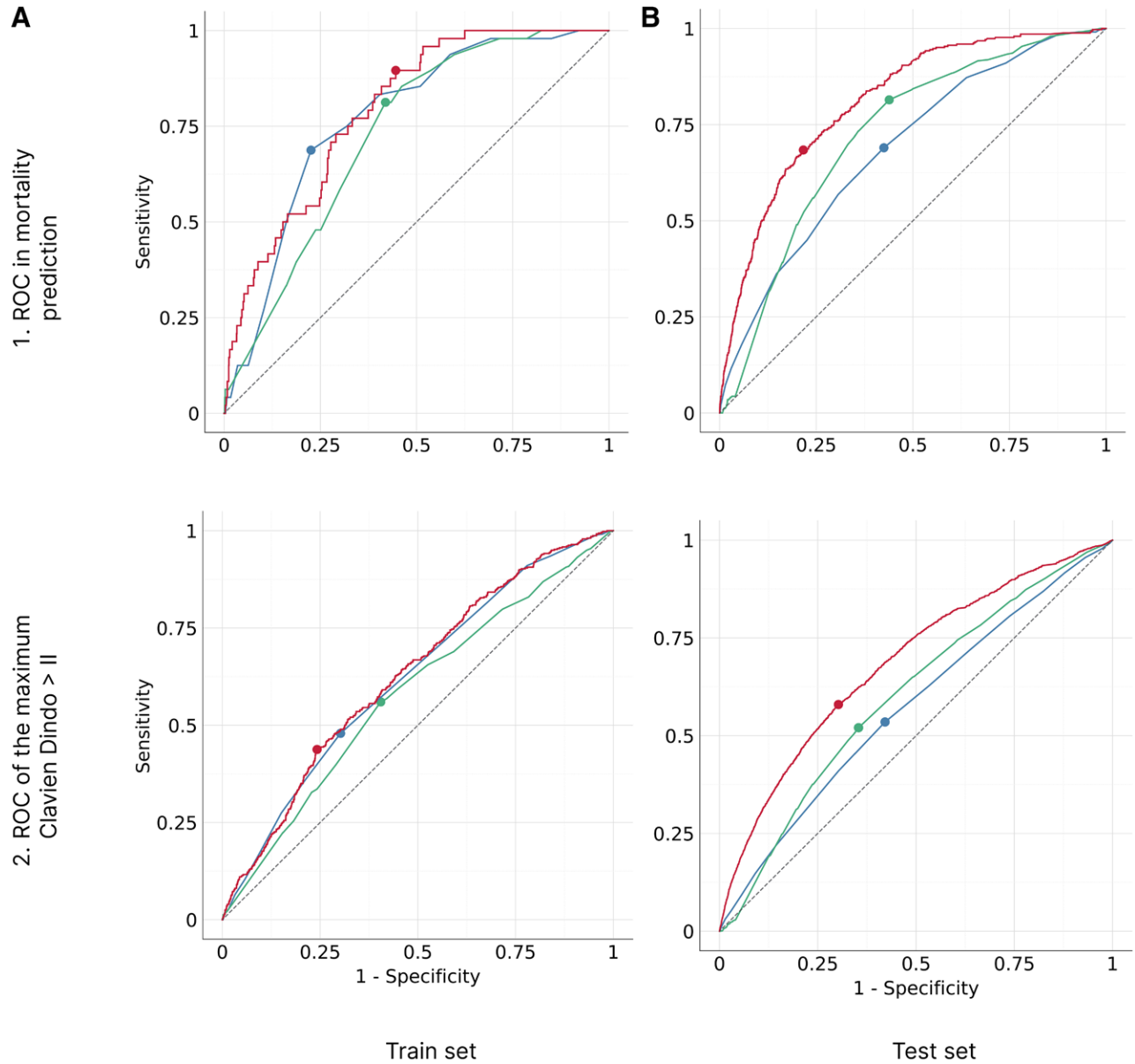

FIGURE 2. Model performances. 1, Receiver operating curve, ROC, of the mortality prediction on (1a) the train set with the area under ROC (AUROC) = 0.77 (0.71; 0.83) for the statistics model (blue), the model with Stabl (red) with AUROC = 0.79 (0.74; 0.85), and the Thoracoscore (green) with AUROC = 0.73 (0.67; 0.79). (1b) The test set with AUROC = 0.69 (0.66; 0.71) for the statistics model (blue), the model with Stabl (red) with AUROC = 0.81 (0.79; 0.83), and the Thoracoscore (green) with AUROC = 0.72 (0.70; 0.75). The gray line represents the random classifier (AUROC = 0.5). 2, ROC of the prediction of the maximum Clavien–Dindo ≥ III on (2a) the train set and (2b) the test set. The AUROC are respectively AUROC = 0.62 (0.59; 0.65) and AUROC = 0.57 (0.56; 0.58) for the statistics model (blue), AUROC = 0.63 (0.60; 0.65) and AUROC = 0.68 (0.68; 0.69). *P* value <1E–16 for the model with Stabl (red) and AUROC = 0.58 (0.55; 0.60) and AUROC = 0.60 (0.59; 0.61) *P* value <1E–16 for the Thoracoscore (green). The dot on each curve represents the probability threshold that optimize the specificity and sensitivity (maximization of Sensitivity–Specificity). The associated metrics are represented in eTable 3 https://links.lww.com/AOSO/A500.

https://links.lww.com/AOSO/A500). Additional performance metrics for prediction are provided in eTable 6, https://links.lww.com/AOSO/A500.

DISCUSSION

The introduction of Predicthor offers a new perspective in cardiothoracic surgical predictive models. While models' predictivity, such as the Thoracoscore's has been validated, Predicthor employs machine learning to extend the capacity of the predictions beyond mortality, capturing also postoperative complications. This stands in contrast to the predominantly mortality-focused approach seen in existing models that also have been more recently questioned for their predictive capacity.

Predicthor is the first model to emphasize data curation as an integral input. Additionally, its dual-layered validation approach, which integrates both internal and external validations, stands out in the field. By harnessing high-quality data from Epithor, this comprehensive evaluation and data preparation strategy not only underlines the model's robustness and reproducibility but also sets it apart from current literature, enhancing Predicthor's credibility and reliability.

Our findings suggest that while Thoracoscore remains a trusted tool, Predicthor, with its interpretable machine learning methodology, delivered superior predictive accuracy in our cohort. Especially when utilizing the Stabl framework, there is a significant improvement over conventional statistical methods, which could be attributed to machine learning's ability to discern

TABLE 2.

Model Equations for Complications and Mortality Predictions

Variable	Statistical model (CCI)	Machine learning model (CCI)	Statistical model (mortality)	Machine learning model (mortality)
Intercept	1.23384167	13.43	1.23384167	-4.94
Age				
<65	0	0	0	0
65–74	+2	1*2.27	+2	1*0.65
75–84	+5	2*2.27	+5	2*0.65
≥ 85	+9	3*2.27	+9	3*0.65
ECOG	+1 per point	0	+1 per point	0
ASA	+1 per point	0	+1 per point	0
Smoking	+3	0	+3	0
FEV1				
Per point	0	-0.08	0	–0.02
≥ 90	0	3*(–0.82)	0	3*0.19
75–89	+3	2*(–0.82)	+3	2*0.19
50–74	+5	1*(–0.82)	+5	1*0.19
<50	+7	0	+7	0
Preoperative tumor stage				
0 or occult	0	0	0	0
IA	0	1*1.01	0	1 × 0.21
IB	0	2*1.01	0	2 × 0.21
IIA	+2	3*1.01	+2	3 × 0.21
IIB	+2	4*1.01	+2	4 × 0.21
IIIA	+3	5*1.01	+3	5 × 0.21
IIIB	+3	6*1.01	+3	6 × 0.21
IV	+3	7*1.01	+3	7 × 0.21
Intervention		0		0
Lobectomie	0		0	
Bilobectomie	+3		+3	
Comorbidity index	0	+0.11 per point	0	+0.08 per point
Comorbidity score	0	–0.4 per point	0	–0.48 per point
Dyspnea	0	+0.84 per point	0	+0.32 per point
Prior thoracic surgery	0		0	
None		0		0
Controlled		1*1.51		1*0.62
Uncontrolled		2*1.51		2*0.62
Metastasis	0		0	
None or Mx		0		0
M0		1*(–2.16)		1*(–1.28)
M1a		2*(–2.16)		2*(–1.28)
M1b		3*(–2.16)		3*(–1.28)
Final computation	Identity	Identity	1/(1+exp(7.5216 – 0.2558*x))	1/(1+exp(–x))

complex relationships that traditional methods might miss or to select new combinations of variables that prove to be more predictive of the surgical outcomes. Additional investigations are required to establish the surgical mortality threshold to be used in conjunction with this model,[23] as per British Thoracic Society guidelines[24,25] recommending that surgical mortality should not exceed 8% for pneumonectomy and 4% for lobectomy.

However, this study presents limitations. The primary dataset, derived from the national EPITHOR database, is retrospective. This, despite its breadth, might introduce inherent biases. Furthermore, the rigorous validation sourced its external validation data from the same database, which could limit its representation of diverse patient populations. The observed overfitting in our models will necessitate further validation against varied, external datasets, including international databases, to ensure the generalizability of Predicthor across different countries. Specifically, studies have shown differences in performance from the Thoracoscore on the British Thoracic Society Database[26] and on the Northern American Thoracic Database. This analysis was not conducted due to insufficient data quality checks on these databases and the inability to calculate the comorbidity index and the comorbidity score. The study also did not include the variable of Vo2max, which is recognized for its predictive strength in assessing postoperative risks following lung resection procedures, due to its absence in the EPITHOR database. This omission is mainly due to the intensive resources required to measure this variable. The limited performance of our models in predicting postoperative complications highlights the need for predictive performance improvement. Incorporating biological signals, such as single-cell and plasma proteomic markers of the host's immune response, could enhance predictive accuracy and better identify patients at risk of complications.[27]

Our findings indicate that Predicthor performs well when applied to segmentectomies and external validation cohorts, suggesting strong potential for broader applicability. While the model was originally trained on a lobectomy-specific cohort, its performance on segmentectomies demonstrates its versatility. To further enhance its predictive accuracy for other surgical procedures, retraining on a dataset that includes a diverse range of cases would be beneficial. This approach would ensure the model continues to deliver reliable and precise predictions across various thoracic surgeries.

In conclusion, Predicthor emerges as a promising tool in cardiothoracic surgery's predictive analytics landscape. Its precision, backed by cutting-edge machine learning, renders it both reliable and comprehensive. While it signifies progress over existing models, continual refinement against diverse datasets is imperative to uphold its relevance and accuracy for a broader range of surgical procedures. Predicthor's potential influence on clinical decisions could enhance patient outcomes and streamline healthcare resource allocation.

ACKNOWLEDGMENTS

X.D. and J.H. contributed equally to this manuscript. J.H. and J.M.B. conceptualized and designed the study. Material preparation and data collection were done by P.A.T., P.B.P., X.B.D., L.B., C.R., P.E.F., and J.M.B. Data analysis were done by X.D. and A.G., and reviewed by J.H. and G.B. All data were accessed and verified by J.M.B. and A.G. The models were developed by X.D. and A.G. The first draft of the manuscript was written by X.D. and J.H. and revised by J.M.B. and A.G. All authors commented on the draft. All authors read and approved the final manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

REFERENCES

1. Weiser TG, Regenbogen SE, Thompson KD, et al. An estimation of the global volume of surgery: a modelling strategy based on available data. *Lancet*. 2008;372:139–144.
2. Boujibar F, Gravier FE, Selim J, et al. Preoperative assessment for minimally invasive lung surgery: need an update? *Thorac Cancer*. 2021;12:3–4.
3. Faujour V, Slim K, Corond P. L'avenir en France de la réhabilitation améliorée après chirurgie, vu sous l'angle médico-économique. *La Presse Médicale*. 2015;44:e23–e31.
4. Falcoz PE, Conti M, Brouchet L, et al. The Thoracic Surgery Scoring System (Thoracoscore): risk model for in-hospital death in 15,183 patients requiring thoracic surgery. *J Thorac Cardiovasc Surg*. 2007;133:325–332.
5. HarpoleDeCampDaley DHMMJ, JrJr, Hur K, Oprian CA, et al. Prognostic models of thirty-day mortality and morbidity after major pulmonary resection. *J Thorac Cardiovasc Surg*. 1999;117:969–979.
6. Yamashita S, Haga Y, Nemoto E, et al. E-PASS (The Estimation of Physiologic Ability and Surgical Stress) scoring system helps the prediction of postoperative morbidity and mortality in thoracic surgery. *Eur Surg Res*. 2004;36:249–255.
7. Kates M, Perez X, Gribetz J, et al. Validation of a model to predict perioperative mortality from lung cancer resection in the elderly. *Am J Respir Crit Care Med*. 2009;179:390–395.
8. O'Dowd EL, Lüchtenborg M, Baldwin DR, et al. Predicting death from surgery for lung cancer: a comparison of two scoring systems in two European countries. *Lung Cancer*. 2016;95:88–93.
9. Chudgar N, Yan S, Hsu M, et al. The American College of surgeons surgical risk calculator performs well for pulmonary resection: a validation study. *J Thorac Cardiovasc Surg*. 2022;163:1509–1516.e1.
10. Lim E, Batchelor T, Shackcloth M, et al; VIOLET Trialists. Study protocol for VIdeo assisted thoracoscopic lobectomy versus conventional Open LobEcTomy for lung cancer, a UK multicentre randomised controlled trial with an internal pilot (the VIOLET study). *BMJ Open*. 2019;9:e029507.
11. Yang CJ, Hartwig MG, D'Amico TA, et al. Large clinical databases for the study of lung cancer: making up for the failure of randomized trials. *J Thorac Cardiovasc Surg*. 2016;151:626–628.
12. Loftus TJ, Tighe PJ, Filiberto AC, et al. Artificial intelligence and surgical decision-making. *JAMA Surg*. 2020;155:148–158.
13. Colborn K, Brat G, Callcut R. Predictive analytics and artificial intelligence in surgery-opportunities and risks. *JAMA Surg*. 2023;158: 337–338.
14. Pforr A, Pagès PB, Baste JM, et al; Epithor Project French Society of Thoracic and Cardiovascular Surgery. A predictive score for bronchopleural fistula established using the French database Epithor. *Ann Thorac Surg*. 2016;101:287–293.
15. Hédou J, Marić I, Bellan G, et al. Discovery of sparse, reliable omic biomarkers with Stabl. *Nat Biotechnol*. 2024;42:1581–1593.
16. Slankamenac K, Graf R, Barkun J, et al. The comprehensive complication index: a novel continuous scale to measure surgical morbidity. *Ann Surg*. 2013;258:1–7.
17. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010;5:1315–1316.
18. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol*. 2010;17:1471–1474.
19. Keung EZ, Gershenwald JE. The eighth edition American Joint Committee on Cancer (AJCC) melanoma staging system: implications for melanoma treatment and care. *Expert Rev Anticancer Ther*. 2018;18:775–784.
20. Erol G, Uzbaş B, Yücelbaş C, et al. Analyzing the effect of data preprocessing techniques using machine learning algorithms on the diagnosis of COVID-19. *Concurr Comput*. 2022;34:e7393.
21. Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann Surg*. 2004;240:205–213.
22. Singh V, Pencina M, Einstein AJ, et al. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Sci Rep*. 2021;11:14490.
23. Sharkey A, Ariyaratnam P, Anikin V, et al. Thoracoscore and European Society objective score fail to predict mortality in the UK. *World J Oncol*. 2015;6:270–275.
24. Dowie J, Wildman M. Choosing the surgical mortality threshold for high risk patients with stage Ia non-small cell lung cancer: insights from decision analysis. *Thorax*. 2002;57:7–10.
25. British Thoracic Society; Society of Cardiothoracic Surgeons of Great Britain and Ireland Working Party. BTS guidelines: guidelines on the selection of patients with lung cancer for surgery. *Thorax*. 2001;56:89–108.
26. Chamogeorgakis T, Toumpoulis I, Tomos P, et al. External validation of the modified Thoracoscore in a new thoracic surgery program: prediction of in-hospital mortality. *Interact Cardiovasc Thorac Surg*. 2009;9:463–466.
27. Rumer KK, Hedou J, Tsai A, et al. Integrated single-cell and plasma proteomic modeling to predict surgical site complications: a prospective cohort study. *Ann Surg*. 2022;275:582–590.