

Failure of Cross-Procedure Transportability of Preoperative Prolonged Length-of-Stay Prediction From VATS to Esophagectomy: A Locked Model Evaluation Study

Michael Mosebetsi Moleka

Renmin Hospital of Wuhan University

Wanli Jiang

Jiang.wanli@qq.com

Renmin Hospital of Wuhan University

Research Article

Keywords: body composition, computed tomography, esophagectomy, external validation, model transportability, postoperative length of stay, prediction model, video-assisted thoracoscopic surgery

Posted Date: September 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10422996/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Prediction models developed in one surgical population may not retain their performance when applied to another procedure, even within the same institution. This study evaluated the cross-procedure transportability of preoperative models for prolonged postoperative length of stay (LOS) developed in patients undergoing video-assisted thoracoscopic surgery (VATS) and applied without modification to patients undergoing esophagectomy.

Methods

This retrospective methodological study included 301 VATS patients and 299 esophagectomy patients treated at Renmin Hospital of Wuhan University. Prolonged LOS was defined according to the recovery context of each procedure as more than 7 days after VATS and more than 15 days after esophagectomy. Two prespecified ridge-logistic models were developed in the VATS cohort: a clinical model and a clinical-plus-computed tomography (CT) model incorporating mean thoracic muscle and fat thickness. Following bootstrap internal validation, the complete VATS prediction pipelines were frozen and evaluated in the esophagectomy cohort without target-cohort tuning. Discrimination, calibration, overall prediction error, prediction distributions, recalibration, and source-to-target case-mix differences were assessed.

Results

Prolonged LOS occurred in 127 VATS patients (42.2%) and 70 esophagectomy patients (23.4%). Optimism-corrected areas under the receiver-operating-characteristic curve (AUROCs) were 0.591 for the clinical model and 0.597 for the clinical-plus-CT model. In the locked esophagectomy evaluation, the corresponding AUROCs were 0.530 (95% confidence interval [CI], 0.457–0.608) and 0.534 (95% CI, 0.456–0.613). The paired AUROC difference was 0.004 (95% interval, – 0.046 to 0.052), providing no evidence that the evaluated CT composites improved ranking performance. Both models overestimated target-cohort risk, with calibration intercepts of – 0.919 and – 0.871. The clinical-plus-CT model generated nearly constant target predictions, rendering its calibration slope nonestimable. Eight of the 10 frozen predictors showed moderate or greater source-to-target imbalance, including five with large standardized differences. Intercept recalibration corrected average-risk differences but did not restore individual-level discrimination.

Conclusions

The VATS-derived prediction pipelines did not transport reliably to the esophagectomy population without modification. These findings define the models' boundary of applicability and demonstrate why locked, procedure-specific evaluation of discrimination, calibration, case mix, and prediction distributions is necessary before clinical implementation.

1. Background

Postoperative length of stay (LOS) is an important indicator of recovery after thoracic surgery, with consequences for patients, caregivers, hospital capacity, and perioperative resource planning. However, LOS is not determined by physiological recovery alone. It also reflects postoperative complications, discharge readiness, clinical pathways, surgeon-level practice, and institutional organization. Substantial variation in LOS has been reported between surgeons following video-assisted thoracoscopic surgery (VATS) lobectomy, even among patients treated within the same healthcare setting [1]. Enhanced-recovery pathways can reduce hospitalization after thoracic surgery, but their implementation and effects remain heterogeneous [2]. Persistent symptoms, air leakage, cardiopulmonary concerns, and logistical barriers may also delay discharge after otherwise uncomplicated VATS [3]. Following esophagectomy, recovery is generally more complex, and considerable hospital-level variation in LOS has been documented [4]. International consensus has therefore emphasized explicit fit-for-discharge criteria rather than reliance on hospitalization duration alone [5]. Standardized esophagectomy pathways may improve perioperative recovery, but differences in pathway implementation and discharge practice remain important [6].

Reliable preoperative prediction of prolonged LOS could support risk communication, perioperative planning, prehabilitation, postoperative surveillance, and early discharge coordination. Recent studies have applied regression and machine-learning methods to predict prolonged hospitalization after VATS, demonstrating the clinical interest in identifying patients at risk of delayed recovery [7]. Machine-learning approaches have also been evaluated for postoperative complications following esophagectomy [8]. Nevertheless, the usefulness of a prediction model depends on more than its apparent performance in the development dataset. Models developed in modest, single-center cohorts may learn associations specific to the local patient population, surgical pathway, predictor distributions, or outcome definition. Their performance may consequently decline when they are applied to a different procedural population, even when the target patients are treated at the same institution.

Computed tomography (CT)-derived body composition may complement conventional preoperative clinical assessment. CT imaging is routinely available for many patients undergoing thoracic oncological surgery, creating an opportunity to extract information about skeletal muscle and adipose tissue without requiring additional testing. Sarcopenia has been associated with adverse short- and long-term outcomes after lung cancer surgery, although published definitions and measurement approaches vary substantially [9]. In patients undergoing esophagectomy, CT-defined sarcopenia and sarcopenic obesity have also been investigated as markers of postoperative complications and survival [10]. Thoracic CT measurements have demonstrated prognostic potential after thoracoscopic pulmonary resection [11],

while both skeletal muscle and mediastinal adipose tissue have been evaluated in surgically treated lung cancer populations [12]. However, recent evidence indicates that muscle-related measurements may provide limited independent information after conventional clinical factors are considered [13]. Biological plausibility should therefore not be assumed to translate automatically into improved prediction. The incremental value of CT-derived variables must be evaluated directly by comparing otherwise identical models with and without the imaging features.

Transportability describes whether a model retains acceptable performance when applied beyond the population and setting in which it was developed. Targeted-validation principles emphasize that a model should be evaluated in the population and clinical context in which it is intended to be used [14]. Generalizability also depends on the relationship between the development and target populations, including differences in case mix, predictor distributions, treatment pathways, and outcome construction [15]. A model may preserve patient ranking while showing poor calibration, or it may fail in both respects; discrimination alone is therefore insufficient to determine whether a model transports successfully [16]. External evaluation should assess discrimination, calibration, overall prediction error, and the distribution of predicted probabilities while maintaining strict separation between model development and target-cohort evaluation [17]. These considerations are particularly important when transferring a prediction pipeline between VATS and esophagectomy, because the procedures differ in operative burden, postoperative management, discharge criteria, baseline risk, and the clinical meaning of prolonged hospitalization.

The present study therefore evaluated whether preoperative prolonged-LOS prediction models developed in a VATS cohort could transport, without modification, to an esophagectomy cohort from the same tertiary institution. Two prespecified ridge-logistic models were examined: a clinical model and a clinical-plus-CT model incorporating mean thoracic muscle and fat thickness. The complete VATS prediction pipelines were frozen before application to the esophagectomy cohort. The primary objective was to quantify cross-procedure transportability through discrimination, calibration, overall prediction error, and prediction-distribution assessment. A further objective was to determine whether the evaluated CT composites provided incremental predictive information beyond routinely available clinical variables. To support interpretation of the locked evaluation, source-to-target differences in the distributions of the frozen predictors were also quantified descriptively. Post-evaluation recalibration and target-data adaptation were examined as exploratory analyses and were not permitted to alter the archived transportability findings.

2. Methods

2.1 Study design, setting, and participants

This retrospective methodological study was conducted at Renmin Hospital of Wuhan University, a tertiary referral center for thoracic surgery. Two independently assembled procedural cohorts were used. Consecutive adult patients who underwent VATS between January 2021 and July 2025 formed the

model-development and internal-validation cohort. Consecutive adult patients who underwent esophagectomy for esophageal carcinoma between December 2020 and June 2025 formed the locked cross-procedure evaluation cohort. Although both cohorts originated from the same institution and were collected during partially overlapping periods, they represented distinct surgical populations with different operative burdens, postoperative pathways, and discharge processes.

Patients were eligible for the VATS cohort if they underwent elective VATS and had preoperative chest CT imaging available within 30 days before surgery, together with the clinical, pulmonary-function, CT-derived, and postoperative LOS information required for the prespecified analysis. Patients were excluded when preoperative CT imaging was unavailable or inadequate for body-composition assessment, when previous thoracic surgery had substantially altered thoracic-wall anatomy, when surgery was performed on an emergency basis, or when a required predictor or outcome value was unavailable.

Patients were eligible for the esophagectomy cohort if they were aged 18 years or older, underwent esophagectomy for esophageal carcinoma, and had preoperative CT imaging suitable for body-composition assessment within 3 months before surgery. Patients undergoing esophagectomy for benign disease or a non-curative or palliative procedure were excluded. Additional exclusions were unavailable or unsuitable CT imaging, incomplete essential clinical or outcome information, early postoperative mortality that prevented interpretation of a conventional discharge-based LOS outcome, and severe concurrent illness—such as end-stage organ dysfunction, dialysis-dependent renal disease, or active systemic infection—that substantially altered the expected postoperative recovery pathway.

After application of the eligibility criteria, 301 VATS patients were included in the development cohort and 299 esophagectomy patients were included in the target cohort. All prespecified predictors and outcomes were complete in the harmonized analytical datasets. The VATS cohort was used exclusively for model development and internal validation. Before row-level esophagectomy outcomes were used for prediction-model evaluation or updating, the complete VATS prediction pipelines—including predictor definitions, preprocessing parameters, penalty values, coefficients, and prediction rules—were frozen and archived. No target-cohort predictor selection, transformation modification, coefficient estimation, penalty tuning, or threshold optimization was permitted before completion of the locked evaluation.

Reporting was guided by the TRIPOD + AI statement [18], and risk-of-bias considerations were structured using PROBAST + AI [19]. The completed TRIPOD + AI checklist and PROBAST + AI self-assessment are supplied as Supplementary Files S1 and S2, respectively.

The study was conducted in accordance with the Declaration of Helsinki and was approved by the institutional review board of Wuhan University Renmin Hospital. The requirement for informed consent was waived due to the retrospective nature of the study.

2.2 Outcomes and predictors

The primary outcome was procedure-specific prolonged postoperative LOS. The recorded postoperative LOS field was used as the authoritative outcome and represented the number of inpatient days from surgery to hospital discharge. Operation and discharge dates were not used to recalculate, correct, or replace the recorded LOS value.

In the VATS development cohort, prolonged LOS was defined as hospitalization for more than 7 days. In the esophagectomy target cohort, prolonged LOS was defined as hospitalization for more than 15 days. These thresholds were prespecified to represent delayed recovery within each procedural context. They were not considered definitions of an identical clinical outcome. The analysis therefore evaluated cross-procedure transportability between related but nonidentical LOS constructs. Postoperative complications, intensive-care requirements, and other events occurring after surgery were not included as predictors because they would not be available during preoperative risk assessment and could introduce outcome leakage.

Candidate predictors were selected a priori according to clinical relevance, routine preoperative availability, and compatibility between the two cohorts. The clinical predictor set included age, sex, BMI, smoking history, hypertension, serum albumin concentration, FEV₁ percentage predicted, and DLCO percentage predicted. Age, BMI, albumin, FEV₁, and DLCO were analyzed as continuous variables. Sex was coded using female as the reference category. Smoking history and hypertension were coded as binary variables, with absence of the characteristic as the reference.

Two primary predictor sets were evaluated. Model 1 contained the eight clinical predictors. Model 2 contained all Model 1 predictors together with mean thoracic muscle thickness and mean thoracic fat thickness. All model variables were available before surgery, preserving the intended use of the models for preoperative risk assessment. The individual T5, T8, and T10 muscle and fat measurements were reserved for a prespecified exploratory sensitivity analysis and were not used to redefine the primary models. Predictor definitions and transformations were frozen before target outcomes were used for model evaluation or updating, consistent with safeguards against model-selection bias and target-data leakage [19].

2.3 CT imaging and body-composition assessment

Preoperative CT examinations obtained during routine clinical evaluation were used for body-composition assessment. In the VATS cohort, the source imaging protocol described multidetector CT acquisition with patients in the supine position, generally during an inspiratory breath hold. Typical acquisition parameters included a tube voltage of 120 kVp, automatic tube-current modulation, and reconstructed slice thicknesses ranging from 1 to 5 mm. The esophagectomy cohort used routine preoperative staging CT obtained within the prespecified 3-month eligibility period. The analytical datasets contained the final CT measurements rather than complete scanner and Digital Imaging and Communications in Medicine acquisition metadata; consequently, scanner-specific parameters were not incorporated into the prediction models.

Body-composition measurements were performed retrospectively on axial CT images at the T5, T8, and T10 vertebral levels using medical-imaging software. Each level was identified using the corresponding vertebral body and adjacent anatomical landmarks. Thoracic muscle thickness was recorded in millimetres as a linear measurement of the posterior thoracic-wall musculature, primarily representing the paraspinal muscle complex and surrounding posterior musculature. Subcutaneous fat thickness was recorded in millimetres from the skin surface to the outer muscular fascia at the corresponding anatomical level. The final analytical variables were linear tissue-thickness measurements rather than cross-sectional areas, muscle-attenuation measurements, or volumetric body-composition estimates.

The harmonized analysis used the same vertebral-level mappings and millimetre units in both cohorts. Only thoracic measurements available in both cohorts were eligible for the cross-procedure analysis. Lumbar measurements available in the esophagectomy source data were not included because compatible lumbar measurements were unavailable in the VATS cohort.

To reduce dimensionality and limit collinearity between anatomically related measurements, two prespecified composite variables were calculated. Mean thoracic muscle thickness was defined as the arithmetic mean of the T5, T8, and T10 muscle measurements. Mean thoracic fat thickness was calculated in the same manner from the corresponding fat measurements. All level-specific measurements were retained in their original millimetre scale before averaging. The three vertebral levels received equal weight, and no sex-specific thresholds, level-specific standardization, outcome-derived cut points, or target-cohort transformations were introduced. Additional CT definitions are provided in Supplementary Methods S1, and the frozen preprocessing parameters are reported in Supplementary Table S1I.

2.4 Data harmonization and model development

Before analysis, the VATS and esophagectomy datasets were harmonized using a prespecified master variable dictionary. Conceptually equivalent variables were assigned common names, units, coding conventions, and reference categories. Continuous clinical and computed tomography (CT) measurements were retained in their original units. Smoking history and hypertension were harmonized as binary variables, and sex was harmonized as female or male. Procedure-specific postoperative length-of-stay (LOS) values and prolonged-LOS indicators were retained separately because the two cohorts used different outcome thresholds. Direct patient identifiers were removed and replaced with cohort-specific study identifiers. The original source workbooks were preserved unchanged.

During data-quality review, one highly sparse VATS spreadsheet row containing only a hypertension entry was identified as a non-patient artifact and removed before creation of the analytical dataset. No otherwise complete clinical record was removed as a duplicate, and every retained analytical record was assigned a unique cohort-specific study identifier. All 299 esophagectomy records were retained. The final harmonized datasets contained 301 VATS records and 299 esophagectomy records. All prespecified predictors and outcomes were complete; therefore, no statistical imputation was required.

Clinical and CT values were not winsorized, manually corrected, or modified on the basis of their distributions or relationships with the outcome.

Two primary ridge-penalized logistic-regression models were developed de novo for the present study using the VATS cohort. Model 1 contained the eight prespecified clinical predictors, whereas Model 2 contained the same clinical predictors together with mean thoracic muscle thickness and mean thoracic fat thickness. Neither model had previously been published. Ridge regression was selected because it permitted simultaneous estimation of the fixed predictor sets while shrinking unstable coefficients and reducing the risk of overfitting in a moderately sized development cohort. The model intercept was not penalized. No univariable screening, stepwise selection, outcome-guided predictor removal, interaction search, nonlinear-term selection, or target-cohort feature selection was performed.

Within each model-fitting sample, continuous predictors were centered and scaled using means and standard deviations estimated exclusively from that sample. Binary predictors retained their prespecified coding. The ridge penalty was selected using stratified five-fold cross-validation within the fitting data, with binomial log loss as the optimization criterion. The one-standard-error rule was used to select the more strongly regularized solution when its cross-validated performance was statistically indistinguishable from that of the minimum-loss penalty.

During internal resampling, preprocessing and penalty selection were repeated within each bootstrap sample to prevent information leakage. For the final frozen models, preprocessing parameters and penalties were estimated using the complete VATS cohort and then applied unchanged to the esophagectomy cohort. No centering, scaling, or penalty parameter was re-estimated using the target data before locked evaluation. In this study, the term “frozen VATS score” refers to the predicted probability generated by one of these locked VATS models and not to a previously published clinical scoring system. Before target-cohort evaluation, the model intercepts, coefficients, penalties, predictor-coding rules, preprocessing parameters, CT-composite definitions, and probability-calculation procedures were fixed and archived. A complete English-language specification of both frozen VATS models is provided in Supplementary File S3 within the Supplementary Appendix. The frozen coefficients, penalties, and preprocessing parameters are additionally reported in Supplementary Tables S1A–S1I.

2.5 Internal validation and locked cross-procedure evaluation

Internal validation was performed in the VATS development cohort using 1,000 full-pipeline bootstrap repetitions. For each repetition, a bootstrap sample was drawn at the patient level, after which centering and scaling parameters were re-estimated, the ridge penalty was retuned by stratified five-fold cross-validation, and the model was refitted. When a patient appeared more than once in a bootstrap sample, all copies of that patient were assigned to the same inner cross-validation fold to prevent duplicate-record leakage.

Model performance was calculated in the bootstrap sample and again when the fitted bootstrap pipeline was applied to the original VATS cohort. The difference between these estimates represented bootstrap optimism. Mean optimism was subtracted from the original apparent performance to obtain the optimism-corrected estimate. The 2.5th and 97.5th percentiles of the replicate-level corrected estimates were reported as 95% empirical optimism intervals. These intervals represented the distribution of bootstrap optimism correction and were not interpreted as conventional independently sampled confidence intervals.

Following internal validation, the final models were fitted using the complete VATS cohort. The resulting preprocessing parameters, penalty values, coefficients, intercepts, analytical code, and prediction rules were stored in a frozen-candidate package with checksum verification. The esophagectomy cohort was not used to select predictors, define transformations, tune source-model penalties, estimate source-model coefficients, or determine the final source-model specification. Locked evaluation began only after the frozen package and its verification manifest had been archived. The complete English-language specification required to reproduce predicted probabilities from the two primary frozen VATS models—including the intended outcome, predictor definitions, coding rules, preprocessing parameters, CT-composite formulas, frozen penalties, model equations, and coefficients—is provided in Supplementary File S3 within the Supplementary Appendix.

For cross-procedure evaluation, each frozen VATS pipeline was applied directly to the 299 esophagectomy patients. Continuous target predictors were transformed using the corresponding VATS means and standard deviations. No target-derived scaling parameters were used. Predicted probabilities were then calculated using the frozen VATS coefficients and intercept. The primary locked evaluation included the clinical ridge model and the clinical-plus-mean-CT ridge model. Neither model was refitted or recalibrated before calculation of the primary target-cohort performance estimates.

Uncertainty in target-cohort performance was assessed using 2,000 stratified bootstrap repetitions that preserved the observed proportion of prolonged-LOS events. The frozen predicted probabilities and observed target outcomes were resampled together, and performance measures were recalculated without model refitting. Percentile-based 95% bootstrap confidence intervals were obtained from the empirical distributions.

Comparisons between Models 1 and 2 were paired. Both models were evaluated in the same bootstrap resample, and the Model 2-minus-Model 1 difference was calculated within every repetition. This preserved within-patient correlation and allowed direct assessment of whether adding the CT composites improved performance beyond the clinical model.

Because prolonged LOS was defined as more than 7 days after VATS and more than 15 days after esophagectomy, the analysis was interpreted as cross-procedure transportability rather than conventional external validation of an identical endpoint. Discrimination assessed whether the frozen VATS scores retained their ability to rank patients in the esophagectomy cohort. Calibration and overall prediction-error measures assessed probability transport across the two procedure-specific outcome

constructs and were interpreted with explicit acknowledgment of endpoint and case-mix differences. All locked-evaluation results were archived before any recalibration or target-cohort adaptation was undertaken.

Complete internal-validation results and paired comparisons are provided in Supplementary Tables S2 and S3 and Supplementary Figure S1. Extended locked-evaluation results, bootstrap-completion diagnostics, and prediction distributions are provided in Supplementary Tables S4–S6 and Supplementary Figure S2.

2.6 Post-evaluation adaptation and learning-curve analysis

All adaptation analyses were performed only after the locked esophagectomy evaluation had been completed, verified, and archived. These analyses were exploratory and were not permitted to modify the frozen VATS models or replace the primary transportability results. Their purpose was to determine which aspects of model performance could potentially be corrected using target-cohort information and to explore how the amount of available esophagectomy data influenced adaptation.

Two conventional recalibration approaches were first examined using the complete esophagectomy cohort. Intercept-only recalibration retained the frozen VATS linear predictor as an offset and estimated a new target-cohort intercept. This adjusted the average predicted risk to the observed esophagectomy event prevalence without changing patient ranking or the relative contribution of individual predictors. Intercept-and-slope recalibration estimated a new intercept and a multiplicative calibration slope for the frozen linear predictor. This approach was considered estimable only when both outcome classes were present, the frozen linear predictor had nonzero variance, and the recalibration model converged. Recalibration based on the complete target cohort was interpreted descriptively because estimation and assessment used the same patients. Complete recalibration parameters are reported in Supplementary Table S7 and Supplementary Figure S3.

The post-evaluation learning-curve analysis used 50 repeated stratified five-fold partitions of the esophagectomy cohort. Within each repetition, each of the five folds served once as the held-out assessment fold, while the remaining observations formed the pool from which adaptation samples were drawn. This produced 250 outer assessment folds for each eligible model, adaptation size, and updating-method combination. Assessment observations were not used for preprocessing, penalty selection, coefficient estimation, or model updating.

Adaptation samples of 50, 100, 150, 200, and 235 patients were drawn while approximately preserving the target-cohort event proportion. These sample sizes corresponded to approximately 12, 23, 35, 47, and 55 prolonged-LOS events, respectively. The same outer partitions were used for both primary models to permit paired interpretation.

Intercept-only recalibration was evaluated at every adaptation size. Intercept-and-slope recalibration was evaluated at adaptation sizes of 100, 150, 200, and 235 patients, subject to the prespecified estimability requirements. Penalized coefficient revision was evaluated only at sample sizes of 200 and 235 because

smaller adaptation samples contained insufficient outcome information for stable multivariable updating.

For coefficient revision, the frozen VATS linear predictor was retained while an unpenalized target intercept and ridge-penalized deviations from the frozen source coefficients were estimated. The revision penalty was selected by stratified inner five-fold cross-validation using only the adaptation sample. The held-out assessment fold was never used for penalty tuning, coefficient estimation, or model selection.

Performance was calculated exclusively in the held-out assessment folds. Learning curves were summarized using mean assessment AUROC. The empirical 2.5th and 97.5th percentiles of repetition-level summaries were reported as stability intervals rather than independent-cohort CIs because the repeated partitions reused observations from the same 299-patient cohort. All nonestimable recalibration attempts, convergence problems, and model-fitting failures were retained and reported rather than replaced with an alternative estimator. Complete learning-curve performance, nonestimability, and parameter results are provided in Supplementary Tables S8–S10. These findings were interpreted as exploratory evidence about adaptation feasibility and data requirements, not as a new independent validation or evidence of clinical readiness.

2.7 Statistical analysis

Model performance was evaluated across discrimination, calibration, and overall prediction accuracy. Discrimination was quantified using the area under the receiver-operating-characteristic curve (AUROC). The area under the precision–recall curve (AUPRC) was also reported because it reflects predictive performance in relation to outcome prevalence. Overall prediction error was assessed using the Brier score and logarithmic loss, for which lower values indicate better performance [16, 17]. The Brier score was a previously established statistical performance measure and was not a prediction model, patient-level clinical score, or measure developed for the present study. It was calculated as the mean squared difference between the observed binary outcome and the predicted probability:

$$\text{Brier score} = (1/n) \sum (y_i - p_i)^2,$$

where y_i represents the observed binary outcome and p_i represents the predicted probability for patient i .

Calibration was evaluated using the calibration intercept, calibration slope, and observed-to-expected (O/E) event ratio. A calibration intercept of zero indicates agreement between average predicted and observed risk, whereas a negative intercept indicates systematic overprediction. A calibration slope of one indicates appropriate spread of predicted risks. The O/E ratio was calculated by dividing the number of observed events by the sum of predicted probabilities; values below one therefore indicate that more events were predicted than observed.

Calibration was additionally examined graphically. Flexible calibration curves were estimated by fitting a restricted cubic spline to the frozen linear predictor, with knots placed at the 10th, 50th, and 90th

percentiles of its distribution. Bootstrap confidence bands were displayed where the curve was estimable. Prediction distributions were examined to determine whether the models generated meaningful separation in estimated risks.

Under the frozen practical-null reporting rule, the calibration slope and flexible calibration curve were classified as nonestimable when the standard deviation of predicted probabilities was 0.001 or less. This rule prevented numerically unstable calibration estimates from being presented as meaningful findings when patients received nearly identical predicted probabilities.

The primary comparison was the clinical-plus-CT ridge model versus the clinical ridge model. Incremental performance was expressed as the paired Model 2-minus-Model 1 difference in AUROC, AUPRC, Brier score, and log loss. For AUROC and AUPRC, positive differences favored Model 2. For the Brier score and log loss, negative differences favored Model 2. Emphasis was placed on effect estimates and uncertainty intervals rather than significance testing. No superiority claim was made when the paired interval included zero.

Prespecified secondary analyses included elastic-net versions of the two fixed predictor sets. The elastic-net mixing parameter was selected from values of 0.10, 0.25, 0.50, 0.75, 0.90, and 1.00 within the inner cross-validation procedure. Sensitivity models consisted of the clinical predictors plus mean muscle thickness alone, the clinical predictors plus mean fat thickness alone, and an exploratory model containing the six individual T5, T8, and T10 muscle and fat measurements. The individual-CT model contained 14 slope parameters and was considered exploratory because of limited development information and anticipated collinearity.

Random forest and gradient-boosting models were not used because the available development sample was insufficient to support reliable tuning of more flexible algorithms. Alternative outcome thresholds and continuous-LOS models were not introduced after examination of the locked results. Published thoracic studies have also used heterogeneous definitions of prolonged LOS, reinforcing the need to interpret the selected thresholds within their procedural contexts [20].

Sample-size adequacy was assessed before model fitting. The VATS cohort contained 127 events, corresponding descriptively to 15.9 events per slope parameter for Model 1 and 12.7 events per slope parameter for Model 2. Formal development-size calculations based on a target shrinkage of 0.90 and a range of plausible model-explanation scenarios indicated that the required sample size exceeded the available 301 patients for both primary models. Ridge penalization was therefore used to reduce coefficient instability but was not considered a substitute for additional outcome information.

The 299-patient esophagectomy cohort contained 70 events. External-validation precision calculations indicated that the expected 95% confidence-interval half-width for the AUROC would be approximately 0.066–0.077 across plausible true AUROCs of 0.65–0.80 [21, 22]. Calibration estimates were expected to be less precise, particularly when the frozen predictions showed limited variation.

Following completion of the primary analysis, source-to-target case-mix differences were quantified descriptively for all 10 frozen predictors. For continuous variables, the standardized mean difference (SMD) was calculated as the target-cohort mean minus the source-cohort mean divided by the square root of the average of the two sample variances. For binary variables, the difference in proportions was divided by the corresponding pooled binomial standard deviation. Absolute SMDs below 0.10 were considered negligible, values of 0.10–0.19 small, values of 0.20–0.49 moderate, and values of 0.50 or greater large. Hypothesis-test p values were not calculated because the purpose was to quantify the magnitude of case-mix shift rather than test whether the cohorts originated from an identical population.

Distributional compatibility was further described using overlap and range-based measures. For continuous variables, distribution overlap was estimated as the integral of the minimum of two Gaussian kernel-density estimates using a common pooled bandwidth; values ranged from zero, indicating no overlap, to one, indicating complete overlap. For binary variables, overlap was calculated as one minus the absolute difference between cohort proportions. Observed-range overlap was calculated as the length of the intersection of the two observed ranges divided by the length of their union. The number and percentage of esophagectomy observations below the VATS minimum or above the VATS maximum were also recorded. This post-evaluation case-mix analysis was conducted solely to improve interpretation of transportability and was not used to alter the predictors, refit either frozen model, or revise the archived performance estimates.

Decision-curve analysis was not performed because no prespecified clinical action, risk threshold, or treatment consequence had been defined for the predicted outcome. Without these elements, net-benefit estimates would not have had a defensible clinical interpretation [23]. The post-evaluation updating hierarchy was limited to intercept recalibration, intercept-and-slope recalibration, and feasibility-gated penalized coefficient revision, consistent with methodological guidance on model updating [24]. Full target-cohort redevelopment was not performed.

All model-development, internal-validation, locked-evaluation, recalibration, and learning-curve analyses were conducted using R version 4.4.2, including glmnet version 4.1–10. The post-evaluation case-mix analysis was performed using a separate reproducible scripted implementation and independently cross-checked. Random seeds, software versions, preprocessing objects, resampling indices, model coefficients, penalties, configuration files, and input and output checksums were retained for reproducibility. Independent quality assurance verified cohort and event counts, CT-composite calculations, predictor mapping, outcome reconstruction, resampling isolation, penalty tuning, frozen-model integrity, and numerical consistency across the final tables and figures. Additional reproducibility information is provided in Supplementary Methods S3.

3. Results

3.1 Cohort characteristics and source-to-target case-mix shift

The VATS development cohort included 301 patients, of whom 127 experienced postoperative LOS longer than 7 days, corresponding to an event prevalence of 42.2%. The locked esophagectomy cohort included 299 patients, of whom 70 experienced postoperative LOS longer than 15 days, corresponding to an event prevalence of 23.4%. No value was missing for any of the 10 predictors included in the two primary models.

The two cohorts differed substantially across several frozen predictors. Eight of the 10 predictors had an absolute SMD of at least 0.20, and five had an absolute SMD of at least 0.50. Compared with the VATS cohort, esophagectomy patients were older, more frequently male, had lower BMI, and had higher mean albumin, FEV₁, DLCO, and thoracic muscle thickness. Mean thoracic fat thickness was substantially lower in the esophagectomy cohort. Smoking history and hypertension were similarly distributed between cohorts (Table 1).

Table 1
Source-to-target comparison of the frozen model predictors

Predictor	VATS development cohort (n = 301)	Esophagectomy target cohort (n = 299)	SMD, target–source	Distribution overlap	Target outside source range
Age, years	60.67 ± 11.25	66.39 ± 7.63	0.596	76.6%	0.0%
Male sex	146 (48.5%)	245 (81.9%)	0.750	66.6%	—
BMI, kg/m ²	23.34 ± 3.45	22.12 ± 2.89	−0.385	83.8%	0.7%
Smoking history	97 (32.2%)	97 (32.4%)	0.005	99.8%	—
Hypertension	91 (30.2%)	86 (28.8%)	−0.032	98.5%	—
Serum albumin, g/L	35.81 ± 3.55	36.84 ± 3.74	0.282	88.1%	2.3%
FEV ₁ , % predicted	82.96 ± 7.40	89.52 ± 3.98	1.105	57.5%	6.7%
DLCO, % predicted	77.95 ± 7.84	85.22 ± 5.10	1.098	59.2%	12.7%
Mean thoracic muscle thickness, mm	25.43 ± 4.29	27.34 ± 4.33	0.441	80.8%	0.3%
Mean thoracic fat thickness, mm	13.91 ± 5.86	8.27 ± 4.02	−1.122	56.5%	4.3%
Notes: Continuous variables are presented as mean ± standard deviation and categorical variables as n (%). The SMD was calculated as the target-cohort value minus the source-cohort value; its sign therefore indicates the direction of the difference. Absolute SMDs of 0.20–0.49 were considered moderate and values of 0.50 or greater large. Distribution overlap ranges from 0%, indicating no overlap, to 100%, indicating complete overlap. “Target outside source range” represents the percentage of esophagectomy observations below the VATS minimum or above the VATS maximum and is not applicable to binary variables. BMI, body mass index; DLCO, diffusing capacity of the lung for carbon monoxide; FEV₁, forced expiratory volume in 1 second; SMD, standardized mean difference; VATS, video-assisted thoracoscopic surgery.					

The largest case-mix differences involved mean thoracic fat thickness, FEV₁, and DLCO, with absolute SMDs of 1.122, 1.105, and 1.098, respectively. These variables also showed the lowest estimated distribution overlap, ranging from 56.5% to 59.2%. In addition, 12.7% of esophagectomy DLCO values and 6.7% of esophagectomy FEV₁ values fell outside the corresponding VATS observed ranges. These findings demonstrated that the development and target populations were not exchangeable with respect to several model predictors.

3.2 VATS model development and internal validation

All 1,000 bootstrap repetitions were completed successfully for each primary model, with no fitting or convergence failure. The apparent AUROC was 0.625 for the clinical model and 0.632 for the clinical-plus-CT model. After bootstrap optimism correction, the corresponding AUROCs were 0.591 and 0.597. Optimism-corrected AUPRC values were 0.525 and 0.542, respectively (Table 2).

Table 2

Apparent and optimism-corrected performance of the primary models in the VATS development cohort

Performance measure	Model 1 apparent	Model 1 corrected (95% interval)	Model 2 apparent	Model 2 corrected (95% interval)
AUROC	0.625	0.591 (0.534–0.644)	0.632	0.597 (0.540–0.646)
AUPRC	0.563	0.525 (0.440–0.611)	0.581	0.542 (0.463–0.627)
Brier score	0.240	0.243 (0.237–0.256)	0.244	0.247 (0.241–0.262)
Log loss	0.674	0.679 (0.668–0.707)	0.681	0.686 (0.675–0.717)
Calibration intercept	0.000	0.001 (– 0.225 to 0.223)	0.000	0.001 (– 0.228 to 0.223)
Calibration slope	5.791	0.654 (– 17.639 to 5.387)	NE	NE
O/E ratio	1.000	1.004 (0.883–1.144)	1.000	1.004 (0.882–1.144)
Notes: Model 1 was the clinical ridge model containing eight predictors. Model 2 was the clinical-plus-mean-CT ridge model containing 10 predictors. Corrected intervals are empirical bootstrap optimism intervals rather than conventional independently sampled CIs. Lower Brier scores and log loss indicate better performance. A calibration intercept of 0, calibration slope of 1, and O/E ratio of 1 represent ideal values. AUPRC, area under the precision–recall curve; AUROC, area under the receiver-operating-characteristic curve; CT, computed tomography; NE, nonestimable; O/E, observed-to-expected.				

The optimism-corrected Model 2-minus-Model 1 difference was 0.006 for AUROC (95% interval, – 0.027 to 0.029) and 0.018 for AUPRC (95% interval, – 0.022 to 0.051). The corresponding differences were 0.004 for the Brier score (95% interval, – 0.001 to 0.010) and 0.008 for log loss (95% interval, – 0.002 to 0.021), with positive values indicating greater prediction error for Model 2. All paired intervals included no difference, providing no convincing evidence that the mean CT composites improved internal performance. Complete results for all seven frozen models are provided in Supplementary Table S2, paired comparisons are provided in Supplementary Table S3, and calibration diagnostics are shown in Supplementary Figure S1.

Calibration estimates were also uncertain. The optimism-corrected calibration slope for Model 1 was 0.654, with a wide empirical interval extending from – 17.639 to 5.387. Model 2 selected a practical-null

ridge solution with insufficient prediction variation for reliable calibration-slope estimation. The CT extension therefore did not produce a stable improvement over the clinical model in the development population.

3.3 Locked cross-procedure evaluation

Both frozen primary models completed all 2,000 target-cohort bootstrap repetitions without execution failure. In the locked esophagectomy cohort, Model 1 produced an AUROC of 0.530 (95% CI, 0.457–0.608), whereas Model 2 produced an AUROC of 0.534 (95% CI, 0.456–0.613). The paired AUROC difference was 0.004 (95% interval, – 0.046 to 0.052). The corresponding AUPRC difference was 0.010 (95% interval, – 0.024 to 0.054). Neither comparison supported improved ranking from adding the mean CT composites (Table 3).

Table 3
Locked performance of the primary VATS models in the esophagectomy cohort

Performance measure	Model 1 estimate (95% CI)	Model 2 estimate (95% CI)	Model 2 – Model 1 difference (95% interval)
AUROC	0.530 (0.457–0.608)	0.534 (0.456–0.613)	0.004 (– 0.046 to 0.052)
AUPRC	0.249 (0.217–0.315)	0.259 (0.219–0.336)	0.010 (– 0.024 to 0.054)
Brier score	0.219 (0.217–0.220)	0.215 (0.215–0.215)	–0.004 (– 0.006 to – 0.003)
Log loss	0.631 (0.627–0.634)	0.622 (0.622–0.622)	–0.009 (– 0.012 to – 0.006)
Calibration intercept	–0.919 (– 0.926 to – 0.912)	–0.871 (– 0.871 to – 0.871)	0.048 (0.042–0.055)
Calibration slope	1.145 (– 2.889 to 5.775)	NE	NE
O/E ratio	0.540 (0.538–0.542)	0.555 (0.555–0.555)	0.015 (0.013–0.017)

Notes: The esophagectomy cohort contained 299 patients and 70 prolonged-LOS events. Model 1 was the clinical ridge model; Model 2 was the clinical-plus-mean-CT ridge model. Positive AUROC and AUPRC differences favor Model 2, whereas negative Brier-score and log-loss differences favor Model 2. Calibration and overall prediction-error measures compare probabilities across different procedure-specific LOS constructs and must not be interpreted as validation of an identical outcome. AUPRC, area under the precision–recall curve; AUROC, area under the receiver-operating-characteristic curve; CI, bootstrap confidence interval; CT, computed tomography; NE, nonestimable; O/E, observed-to-expected; VATS, video-assisted thoracoscopic surgery.

The lower Brier score and log loss for Model 2 reflected its narrow, nearly constant prediction distribution rather than improved individual-level discrimination. The mean predicted probability was 0.434 for Model 1 and 0.422 for Model 2, substantially higher than the observed esophagectomy event

prevalence of 0.234. The corresponding calibration intercepts were – 0.919 and – 0.871, while the O/E ratios were 0.540 and 0.555. Both models therefore systematically overestimated the target procedure-specific event rate.

Model 1 had a calibration slope of 1.145, but its 95% CI was extremely wide (– 2.889 to 5.775), indicating limited precision. Model 2 had a target prediction standard deviation of 0.00036, below the prespecified practical-null threshold. Its calibration slope and flexible calibration curve were consequently nonestimable. The calibration display therefore showed an estimable but restricted curve for Model 1 and explicitly reported nonestimability for Model 2 (Fig. 1).

Model 1 predictions ranged from 0.392 to 0.481, with a standard deviation of 0.0150. Model 2 predictions ranged from 0.4210 to 0.4232, with a standard deviation of 0.00036. Event and non-event prediction distributions overlapped extensively for both models, particularly for Model 2 (Fig. 2 and Supplementary Table S6).

Extended performance results for all seven frozen pipelines and complete bootstrap-completion and estimability results are provided in Supplementary Tables S4 and S5 and Supplementary Figure S2.

3.4 Post-evaluation recalibration

Intercept-only recalibration estimated target-cohort intercept adjustments of – 0.919 for Model 1 and – 0.871 for Model 2. These adjustments aligned the mean predicted probabilities with the observed esophagectomy event prevalence and reduced the descriptive Brier scores to 0.1792 and 0.1793, respectively. The corresponding log losses were 0.5438 and 0.5442. Because intercept recalibration applies the same log-odds adjustment to every patient, it did not change patient ranking or AUROC (Table 4).

Table 4
Descriptive full-cohort recalibration of the primary models

Model	Recalibration method	Intercept (95% CI)	Slope (95% CI)	Brier score	Log loss
Model 1	Intercept only	–0.919 (– 0.926 to – 0.912)	1.000, fixed	0.1792	0.5438
Model 1	Intercept and slope	–0.881 (– 1.966 to 0.334)	1.145 (– 2.889 to 5.775)	0.1792	0.5438
Model 2	Intercept only	–0.871 (– 0.871 to – 0.871)	1.000, fixed	0.1793	0.5442
Model 2	Intercept and slope	NE	NE	NE	NE

Notes: These are descriptive apparent estimates because recalibration and performance assessment used the same 299-patient esophagectomy cohort. Intercept-only recalibration changes average predicted risk but does not change ranking. Model 2 intercept-and-slope recalibration was nonestimable under the frozen practical-null prediction-variation rule. CI, bootstrap confidence interval; NE, nonestimable.

Intercept-and-slope recalibration remained highly uncertain for Model 1. For Model 2, it was not estimable because the frozen linear predictor contained insufficient variation. Thus, recalibration corrected the average-risk offset but did not establish reliable individualized prediction or recover meaningful ranking performance. Complete parameters are provided in Supplementary Table S7 and Supplementary Figure S3.

3.5 Exploratory target-data learning curve

All 50 repeated stratified five-fold partitions were completed. Intercept-only recalibration was estimable at every adaptation size for both models and produced assessment AUROCs that remained essentially unchanged from the frozen models. Model 1 intercept-and-slope recalibration was unstable and did not improve assessment performance. Model 2 intercept-and-slope recalibration was nonestimable in every eligible repetition because the frozen predictions remained practically constant.

Ridge coefficient revision was estimable in all 50 repetitions at adaptation sizes of 200 and 235 patients. At an adaptation size of 235, the mean assessment AUROC was 0.538 for Model 1, with an empirical stability interval of 0.513–0.569. For Model 2, the corresponding mean AUROC was 0.575, with a stability interval of 0.524–0.617. Mean Brier scores and log losses changed little with increasing adaptation size (Table 5).

Table 5
Exploratory performance after ridge coefficient revision

Frozen model	Adaptation sample	AUROC, mean (stability interval)	AUPRC, mean (stability interval)	Brier score, mean (stability interval)	Log loss, mean (stability interval)	Usable repetitions
Model 1	200	0.536 (0.519–0.567)	0.251 (0.236–0.277)	0.1791 (0.1787–0.1800)	0.5437 (0.5422–0.5460)	50/50
Model 1	235	0.538 (0.513–0.569)	0.249 (0.232–0.270)	0.1791 (0.1787–0.1797)	0.5437 (0.5423–0.5453)	50/50
Model 2	200	0.570 (0.514–0.610)	0.275 (0.245–0.312)	0.1793 (0.1787–0.1802)	0.5442 (0.5424–0.5465)	50/50
Model 2	235	0.575 (0.524–0.617)	0.274 (0.242–0.309)	0.1793 (0.1786–0.1802)	0.5440 (0.5422–0.5463)	50/50

Notes: Values are means across 50 repeated stratified five-fold partitions. Stability intervals are the empirical 2.5th and 97.5th percentiles of repetition-level summaries. They are not independent-cohort CIs because the same 299 patients were reused across overlapping partitions. Ridge coefficient revision estimated penalized deviations from the frozen VATS coefficients using only the adaptation sample. AUPRC, area under the precision–recall curve; AUROC, area under the receiver–operating-characteristic curve.

The learning curves showed modest improvement in mean AUROC after coefficient revision, particularly for Model 2, but substantial uncertainty remained (Fig. 3). These exploratory estimates arose from repeated partitions of a single 299-patient cohort with 70 events and did not constitute independent external validation. Results for every adaptation size and updating method, together with failure and parameter summaries, are provided in Supplementary Tables S8–S10.

3.6 Secondary analyses and quality assurance

The secondary elastic-net clinical and clinical-plus-CT models produced locked esophagectomy AUROCs of 0.583 (95% CI, 0.511–0.658) and 0.584 (95% CI, 0.510–0.660), respectively. The muscle-only, fat-only, and individual-CT ridge sensitivity models produced AUROCs of 0.539, 0.523, and 0.550, respectively. None of the secondary or sensitivity analyses demonstrated a convincing incremental ranking advantage from adding the evaluated CT measurements. Complete internal- and target-cohort performance results for all seven frozen pipelines are provided in Supplementary Tables S2 and S4.

All 7,000 planned internal-validation fits and all 14,000 locked target-cohort bootstrap evaluations across the seven frozen pipelines completed without execution failure. Nonestimable calibration quantities reflected insufficient prediction variation rather than model-fitting failure. Independent quality assurance confirmed the reconstruction of cohort and event counts, CT composites, predictor mappings, preprocessing isolation, inner-fold grouping, penalty tuning, frozen coefficients, paired metric calculations, deterministic seeds, and consistency of the manuscript tables and figures. The post-evaluation case-mix calculations were independently cross-checked. The original workbooks, harmonized datasets, and frozen VATS prediction package remained unchanged throughout the locked evaluation and subsequent manuscript analyses.

4. Discussion

The principal contribution of this study is a locked demonstration that a prediction pipeline developed in one thoracic surgical population cannot be assumed to remain valid when transferred to another procedure, even when both populations originate from the same institution. The clinical and clinical-plus-CT models showed only modest discrimination after optimism correction in the VATS cohort and near-chance discrimination when applied unchanged to patients undergoing esophagectomy. Adding mean thoracic muscle and fat thickness did not produce a convincing improvement in ranking in either cohort. The paired target-cohort AUROC difference was only 0.004, with an interval compatible with both modest benefit and modest harm. Both frozen models also substantially overestimated the target procedure-specific event rate, as shown by negative calibration intercepts and O/E ratios close to 0.55. The clinical-plus-CT model generated almost constant esophagectomy predictions, preventing meaningful estimation of its calibration slope or flexible calibration curve. Importantly, these findings should not be reduced to the statement that the models performed poorly. The study successfully executed a prespecified transportability experiment and identified the boundary beyond which the frozen prediction functions no longer retained adequate performance. This is precisely the purpose of targeted validation:

to determine whether a model is suitable for the population and clinical setting in which it might actually be used [14]. The results also support the broader principle that model performance is conditional on the joint distribution of predictors, outcome, population, and healthcare context rather than being an intrinsic property of an algorithm [15].

Several connected mechanisms probably contributed to the observed loss of transportability, although the present design cannot isolate the independent effect of any single mechanism. First, the model crossed a major procedural boundary. Recovery after VATS is strongly influenced by pulmonary status, air leakage, pain control, and surgeon- or pathway-specific discharge practice [1]. Esophagectomy involves a different physiological burden and recovery process, including nutritional management, anastomotic surveillance, and more complex fit-for-discharge considerations [5]. Second, the outcome itself shifted. VATS LOS longer than 7 days and esophagectomy LOS longer than 15 days are clinically related indicators of delayed recovery but are not identical endpoints. The resulting differences in baseline prevalence and clinical meaning directly affect calibration, Brier score, log loss, and AUPRC. Third, the case-mix analysis demonstrated that the source and target populations were not exchangeable. Eight of the 10 frozen predictors showed moderate or greater imbalance, and five showed large imbalance. The greatest shifts occurred in thoracic fat thickness, FEV₁, DLCO, sex, and age. Distribution overlap was particularly limited for thoracic fat and pulmonary-function measures, and a meaningful proportion of target DLCO values fell outside the source range. These differences provide direct evidence that several model inputs occupied different regions of the predictor space. Nevertheless, procedure, case mix, outcome construction, imaging measurement, and care pathway changed together; the decline in performance cannot be attributed exclusively to procedure type. The CT findings also require appropriately specific interpretation. The study shows that these raw, equally weighted T5, T8, and T10 thickness composites did not provide convincing incremental value in these models. It does not establish that CT-derived body composition is generally uninformative. Prior studies have reported associations between sarcopenia and outcomes after lung cancer surgery [9], esophagectomy [10], and thoracoscopic pulmonary resection [11]. More recent evidence continues to associate sarcopenia with outcomes after surgically treated non-small cell lung cancer [25], while frailty has been associated with complications after esophagectomy [26]. These findings support continued evaluation of physiological reserve but do not establish incremental predictive value for a particular imaging measurement. Alternative approaches—such as skeletal-muscle area, muscle attenuation, volumetric measurements, sex-specific indices, or formally validated repeat-reader protocols—may capture physiological reserve more effectively than the linear composites evaluated here.

The post-evaluation analyses clarify what model updating could and could not accomplish. Intercept-only recalibration corrected the difference between average predicted risk and the target event prevalence, reducing the Brier score and log loss, but it left patient ordering unchanged. This is expected because an intercept adjustment applies the same log-odds shift to every prediction. The improvement in aggregate prediction error therefore represented correction of average-risk offset rather than restoration of individualized prediction. Model 1 intercept-and-slope recalibration was highly uncertain, while Model 2 slope recalibration was nonestimable because its predictions contained almost no

variation. These findings illustrate why recalibration is most useful when a model preserves meaningful ranking but systematically misestimates baseline risk; it cannot create discrimination that is absent from the frozen score [24]. Ridge coefficient revision produced modestly higher mean AUROC values at the largest adaptation size, reaching 0.538 for Model 1 and 0.575 for Model 2. However, these estimates remained uncertain and were derived from overlapping partitions of the same 299 patients and 70 events. They do not overturn the locked evaluation and should not be interpreted as independent validation of revised models. Clinically, neither frozen model nor its exploratory adaptations should currently guide individual counseling, resource allocation, or discharge planning. The appropriate next step is not simply to apply a more complex algorithm. Future work should define the intended clinical action, align the outcome with that action, establish standardized discharge criteria, and recruit a larger procedure-specific development cohort. Evaluation should report calibration and overall accuracy alongside discrimination, as recommended for external-validation studies [16, 17]. Decision-curve analysis would become appropriate only after a defensible action and risk-threshold range have been specified [23]. Ultimately, any proposed model-guided pathway should be tested prospectively, because predictive accuracy alone does not demonstrate that using a model improves care [27].

This study has several strengths that support the credibility of its central conclusion. Predictor sets, preprocessing, model classes, penalty-selection procedures, evaluation metrics, and adaptation rules were specified before model fitting. Row-level target outcomes were not used to select or tune the VATS models, and the complete source pipelines were frozen before locked evaluation. Preprocessing and tuning were repeated within resampling, paired bootstraps were used for incremental comparisons, and the target evaluation was archived before recalibration. The study examined discrimination, calibration, overall prediction error, prediction distributions, estimability, and case-mix shift rather than relying on AUROC alone. The original workbooks, harmonized datasets, configuration files, coefficients, and analytical artifacts were checksum-protected, and independent quality assurance verified the principal numerical results. These features are consistent with the transparency expected under TRIPOD + AI [18] and address several risks emphasized by PROBAST + AI [19]. The study nevertheless remains retrospective and single-center, and the development and target cohorts were of moderate size. Formal development-size criteria were not met, and the 70 target events limited the precision of calibration estimates, consistent with contemporary external-validation sample-size guidance [21, 22]. The procedure-specific LOS definitions prevented interpretation as conventional external validation of an identical outcome, although this construct shift was also an intentional part of the transportability question. CT predictors were based on linear thickness rather than standardized area, attenuation, or volumetric measurements, and repeat-reader reliability estimates were unavailable in the analytical records. The study also lacked an independent contemporary VATS cohort, so it does not establish temporal or geographic transportability within the source procedure. These limitations narrow the scope of inference but do not invalidate the primary observation: when a fully frozen prediction pipeline was transferred across a clinically important procedural boundary, its discrimination, calibration, and prediction distribution were inadequate for clinical use. The study therefore provides positive

methodological evidence for rigorous procedure-specific validation before implementation rather than merely documenting an unsuccessful model.

5. Conclusions

This locked cross-procedure evaluation demonstrated that preoperative prolonged-LOS prediction models developed in a VATS population did not transport reliably to patients undergoing esophagectomy without modification. The frozen models showed near-chance discrimination, substantial average-risk overprediction, and limited separation of individual predicted risks. Adding the evaluated mean thoracic muscle and fat thickness composites did not provide convincing incremental ranking value. Although intercept recalibration corrected the average-risk difference, it could not restore patient-level discrimination, and exploratory coefficient revision remained uncertain. These findings define the boundary of applicability of the evaluated models and demonstrate why prediction pipelines require procedure-specific assessment of case mix, discrimination, calibration, prediction distributions, and estimability before clinical implementation. Future models should be developed in larger, procedure-aligned cohorts using clinically harmonized outcomes, standardized imaging measurements, and independent validation before prospective evaluation of their effect on patient care.

Abbreviations

AUPRC, area under the precision–recall curve

AUROC, area under the receiver-operating-characteristic curve

O/E, observed-to-expected event ratio

PROBAST+AI, Prediction Model Risk of Bias Assessment Tool plus Artificial Intelligence

SMD, standardized mean difference

TRIPOD+AI, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis plus Artificial Intelligence

VATS, video-assisted thoracoscopic surgery

Declarations

Ethics approval and consent to participate

This retrospective study was approved by the Institutional Review Board/Ethics Committee of Renmin Hospital of Wuhan University. The requirement for informed consent was waived because of the retrospective nature of the study and the use of de-identified clinical data.

Clinical trial number

Clinical trial number: not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The de-identified participant-level data underlying this study are not publicly available because they contain confidential clinical information. They may be made available by the corresponding author on reasonable request, subject to approval by the Institutional Review Board of Renmin Hospital of Wuhan University and compliance with applicable institutional and data-protection requirements. The frozen model specifications, aggregate results, and reporting materials are provided in the article and its supplementary files.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors received no specific funding for this work.

Authors' contributions

MMM contributed to the study conception and design, data analysis and interpretation, drafting of the manuscript, and critical revision of the manuscript. WJ contributed to study supervision, clinical interpretation of the findings, and critical revision of the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Authors' information

Michael Mosebetsi Moleka and Wanli Jiang are affiliated with the Department of Cardiothoracic Surgery, Renmin Hospital of Wuhan University, Wuhan, China.

References

1. Zini J, Dayan G, Têtu M, Kfoury T, Maqueda LB, Abdulnour E, et al. Intersurgeon variations in postoperative length of stay after video-assisted thoracoscopic surgery lobectomy. *JTCVS Open*. 2024;18:253–60. 10.1016/j.xjon.2024.01.003. . PMID:38690406.

2. Khoury AL, McGinagle KL, Freeman NL, El-Zaatari H, Feltner C, Long JM, University of North Carolina School of Medicine Enhanced Recovery Program Working Group. Enhanced recovery after thoracic surgery: systematic review and meta-analysis. *JTCVS Open*. 2021;7:370–91. 10.1016/j.xjon.2021.07.007. . PMID:36003715.
3. Huang L, Kehlet H, Petersen RH. Reasons for staying in hospital after video-assisted thoracoscopic surgery lobectomy. *BJS Open*. 2022;6(3). 10.1093/bjsopen/zrac050. . PMID:35511502.
4. Voeten DM, van der Werf LR, van Sandick JW, van Hillegersberg R, van Berge Henegouwen MI, Dutch Upper Gastrointestinal Cancer Audit Group. Length of hospital stay after uncomplicated esophagectomy: hospital variation shows room for nationwide improvement. *Surg Endosc*. 2021;35(11):6344–57. 10.1007/s00464-020-08103-4. . PMID:33104919.
5. Muller PC, Kapp JR, Vetter D, Gutschow CA, Kappeler A, Weitz J, et al. Fit-for-discharge criteria after esophagectomy: an international expert Delphi consensus. *Dis Esophagus*. 2021;34(6). 10.1093/dote/daaa101. . PMID:32960264.
6. Puccetti F, Wijnhoven BPL, Kuppusamy M, Hubka M, Low DE. Impact of standardized clinical pathways on esophagectomy: a systematic review and meta-analysis. *Dis Esophagus*. 2022;35(2). 10.1093/dote/daab027. . PMID:34009322.
7. Zhang G, Liu X, Hu Y, Luo Q, Ruan L, Xie H et al. Development and comparison of machine-learning models for predicting prolonged postoperative length of stay in lung cancer patients following video-assisted thoracoscopic surgery. *Asia Pac J Oncol Nurs*. 2024;11(6):100493. 10.1016/j.apjon.2024.100493. PMID:38808011.
8. Jung JO, Pisula JI, Bozek K, Popp F, Fuchs HF, Schröder W, et al. Prediction of postoperative complications after oesophagectomy using machine-learning methods. *Br J Surg*. 2023;110(10):1361–6. 10.1093/bjs/znad181. . PMID:37343072.
9. Kawaguchi Y, Hanaoka J, Ohshio Y, Okamoto K, Kaku R, Hayashi K, et al. Does sarcopenia affect postoperative short- and long-term outcomes in patients with lung cancer? A systematic review and meta-analysis. *J Thorac Dis*. 2021;13(3):1358–69. 10.21037/jtd-20-3072. . PMID:33841929.
10. Fehrenbach U, Wuensch T, Gabriel P, Segger L, Yamaguchi T, Auer TA, et al. CT body composition of sarcopenia and sarcopenic obesity: predictors of postoperative complications and survival in patients with locally advanced esophageal adenocarcinoma. *Cancers (Basel)*. 2021;13(12):2921. 10.3390/cancers13122921. . PMID:34208070.
11. Hasenauer A, Forster C, Hungerbühler J, Perentes JY, Abdelnour-Berchtold E, Koerfer J, et al. CT-derived sarcopenia and outcomes after thoracoscopic pulmonary resection for non-small cell lung cancer. *Cancers (Basel)*. 2023;15(3):790. 10.3390/cancers15030790. . PMID:36765748.
12. Ponholzer F, Groemer G, Ng C, Maier H, Lucciarini P, Kocher F, et al. Sarcopenia and mediastinal adipose tissue as a prognostic marker for short- and long-term outcomes after primary surgical treatment for lung cancer. *Cancers (Basel)*. 2023;15(23):5666. 10.3390/cancers15235666. . PMID:38067372.

13. Han DJ, Na KJ, Yun T, Park JH, Na B, Park S, et al. Effects of respiratory sarcopenia on the postoperative course in elderly lung cancer patients: a retrospective study. *J Cardiothorac Surg*. 2025;20:71. 10.1186/s13019-024-03185-w. . PMID:39827359.
14. Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. *Diagn Progn Res*. 2022;6(1):24. 10.1186/s41512-022-00136-8. . PMID:36550534.
15. Wan B, Caffo B, Vedula SS. A unified framework on generalizability of clinical prediction models. *Front Artif Intell*. 2022;5:872720. 10.3389/frai.2022.872720. PMID:35573904.
16. Collins GS, Dhiman P, Ma J, Schlüssel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*. 2024;384. 10.1136/bmj-2023-074819. . PMID:38191193.
17. 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;384. 10.1136/bmj-2023-074820. . PMID:38224968.
18. 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. 10.1136/bmj-2023-078378. . PMID:38626948.
19. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST + AI: an updated quality, risk-of-bias, and applicability assessment tool for prediction models using regression or artificial-intelligence methods. *BMJ*. 2025;388. 10.1136/bmj-2024-082505. . PMID:40127903.
20. Zheng YL, Huang RS, Liang XY. Identification of preoperative risk factors associated with prolonged length of stay after lobectomy. *Heliyon*. 2023;9(5). 10.1016/j.heliyon.2023.e16061. . PMID:37206051.
21. Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, Van Calster B, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *BMJ*. 2024;384. 10.1136/bmj-2023-074821. . PMID:38253388.
22. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med*. 2021;40(19):4230–51. 10.1002/sim.9025. . PMID:34031906.
23. Piovani D, Sokou R, Tsantes AG, Vitello AS, Bonovas S. Optimizing clinical decision making with decision curve analysis: insights for clinical investigators. *Healthc (Basel)*. 2023;11(16):2244. 10.3390/healthcare11162244. . PMID:37628442.
24. Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. *BMC Med Res Methodol*. 2022;22(1):316. 10.1186/s12874-022-01801-8. . PMID:36510134.
25. Lading T, Media AS, Tellefsen B, Hansen SN, Schmidt JA, Christensen TD et al. Sarcopenia and survival in surgically treated non-small cell lung cancer: a systematic review and meta-analysis. *Eur*

J Surg Oncol. 2025;51(12):110538. 10.1016/j.ejso.2025.110538. PMID:41138294.

26. Kamińska A, Bonczar M, Plutecki D, Ostrowski P, Koziej M, Konturek A. Postoperative complications in patients with frailty syndrome undergoing esophagectomy: a systematic review with meta-analysis. J Clin Med. 2026;15(8):3040. 10.3390/jcm15083040. . PMID:42074841.
27. Chen X, Xi Y, Lin C, Yu H, Kim JY, Wee JO, et al. An esophagectomy surgical Apgar score-based nomogram—a risk-based postoperative triage system: a phase II randomised trial. J Thorac Dis. 2024;16(4):2550–62. 10.21037/jtd-24-451. . PMID:38738231.

Figures

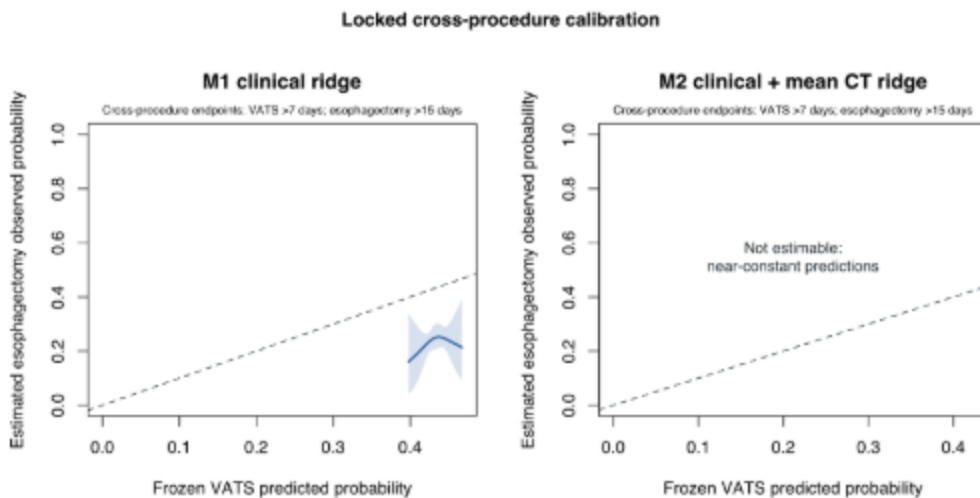


Figure 1

Locked cross-procedure calibration of the primary VATS models in the esophagectomy cohort. The dashed diagonal represents perfect agreement between predicted and observed probabilities. The Model 1 curve is displayed only over the supported range of frozen predictions, with its bootstrap confidence band. Model 2 calibration was nonestimable because the frozen probabilities were nearly constant. The figure compares probabilities across procedure-specific endpoints and should not be interpreted as calibration for an identical outcome.

Locked prediction distributions in esophagectomy

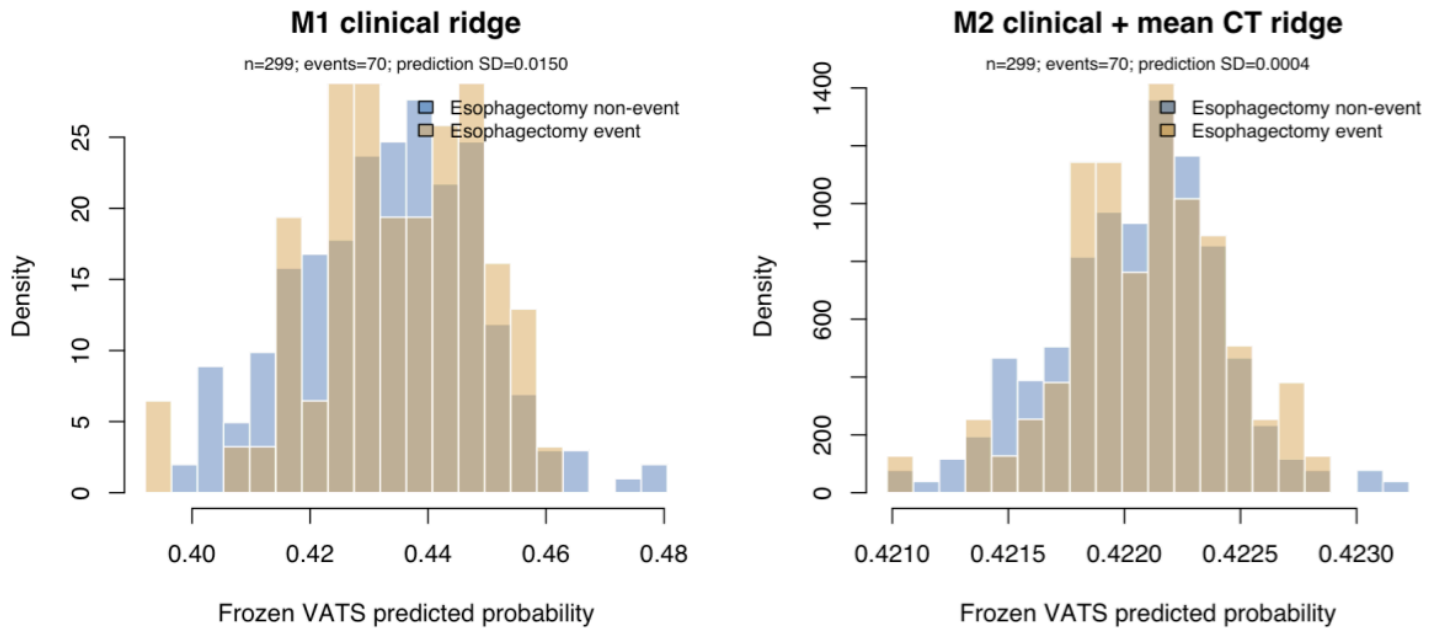


Figure 2

Distribution of frozen predicted probabilities in the esophagectomy cohort. Histograms show the distributions of predicted probabilities among patients with and without procedure-specific prolonged LOS. The left panel presents Model 1, and the right panel presents Model 2. Extensive overlap indicates limited patient-level separation; the different horizontal scales reflect the substantially narrower prediction range of Model 2.

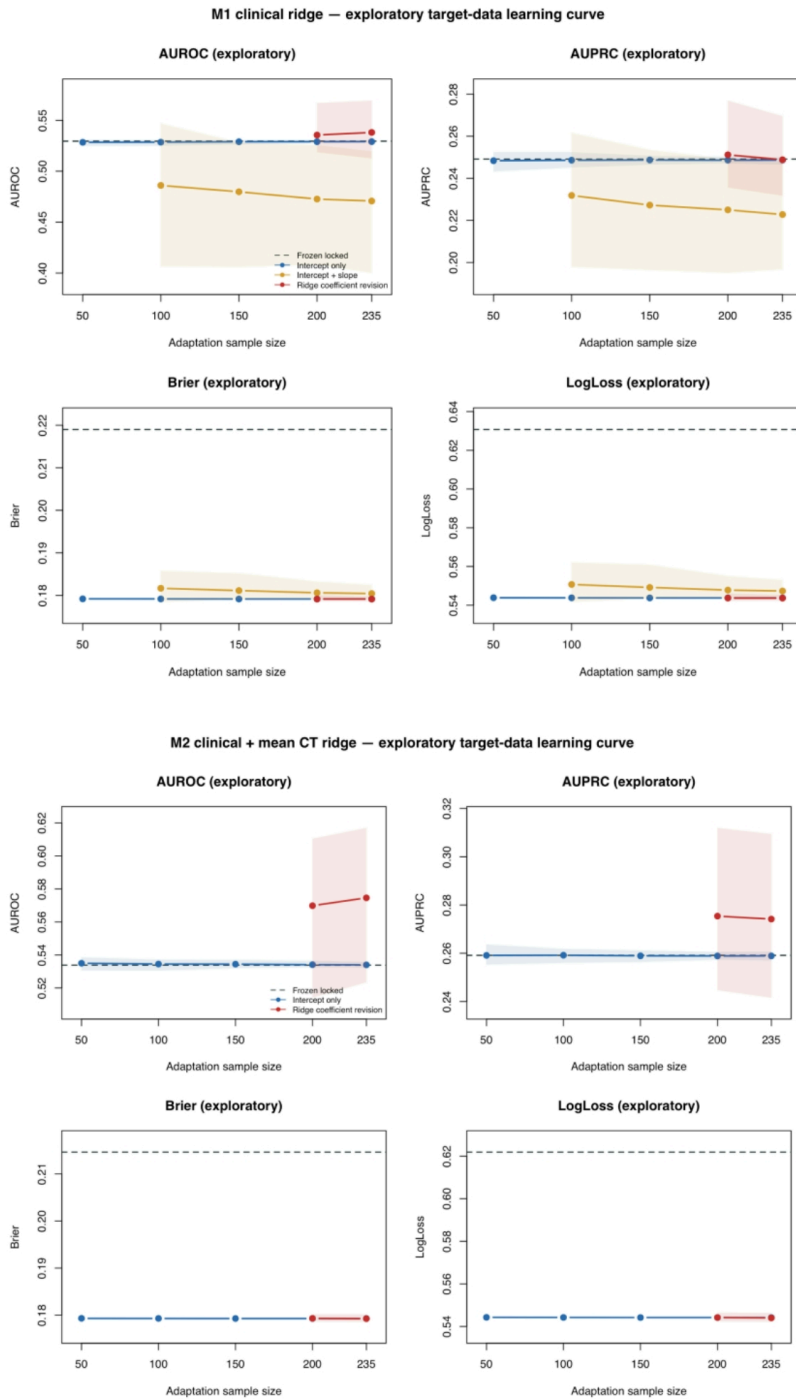


Figure 3

Exploratory target-data learning curves. Panel A shows results for Model 1, and Panel B shows results for Model 2. Points represent mean performance across 50 repeated stratified five-fold partitions, and shaded regions represent empirical stability intervals. The dashed line represents performance of the archived frozen model. Intercept-only recalibration, intercept-and-slope recalibration where estimable,

and ridge coefficient revision are shown. Model 2 intercept-and-slope recalibration was nonestimable at all planned adaptation sizes.

Supplementary Files

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

- [VATStoESOSupplementaryAppendix.docx](#)