SRIWIJAYA JOURNAL OF SURGERY
Vol. 9 No. 2 (2026), pp. 254-274 · sriwijayasurgery.com
e-ISSN 2722-3558

ORIGINAL ARTICLE

Thoracoscore Risk Bands Substantially Underpredict In-Hospital Mortality After Thoracic Surgery: An External Validation Study in a Tertiary Referral Cohort

Alif Alfiansyah[1]*, Gama Satria[2], Theodorus[3]

[1] Department of Surgery, Faculty of Medicine, Universitas Sriwijaya, Palembang, Indonesia
[2] Department of Thoracic, Cardiac, and Vascular Surgery, Faculty of Medicine, Universitas Sriwijaya, Palembang, Indonesia
[3] Department of Pharmacology, Faculty of Medicine, Universitas Sriwijaya, Palembang, Indonesia

ARTICLE INFO

Article history:
Received: August 12, 2026
Revised: September 15, 2026
Accepted: October 10, 2026
Available Online: October 20, 2026
Published: October 28, 2026

Keywords:
Calibration; External validation; Surgical mortality; Thoracic surgery; Thoracoscore

Corresponding author:
Alif Alfiansyah
Email:
alif.alfiansyah1992@gmail.com

DOI:
https://doi.org/10.37275/sjs.v9i2.157

Access licence:
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0).

Copyright:
© 2026 The Authors. Authors retain copyright and grant the journal the right of first publication. Published by HM Publisher.

ABSTRACT

Background: Thoracoscore is widely used to estimate preoperative mortality after general thoracic surgery, but its performance in Southeast Asian referral practice is uncertain.

Objective: To evaluate the discrimination and calibration of the Thoracoscore risk-band table used routinely at an Indonesian tertiary centre.

Methods: This retrospective cohort included 106 consecutive adults who underwent thoracic surgery at Dr Mohammad Hoesin General Hospital, Palembang, from January to December 2025. Seven of nine Thoracoscore domains were retrievable; therefore, the published points-to-risk band table, rather than the original logistic equation, was validated. Discrimination was assessed by permutation testing, calibration by observed-to-expected (O:E) ratios and within-band tests, and independent predictors by Firth-penalised logistic regression.

Result: In-hospital mortality was 26.4% (28/106; 95% CI 19.0–35.5). Discrimination was absent (AUC 0.572, 95% CI 0.457–0.687; permutation p = 0.249). Calibration failed in every risk band: 28 deaths occurred compared with 2.88 expected (O:E 9.71, 95% CI 6.45–14.03), with a minimum O:E of 4.89 under the most favourable assignment of the two unrecorded domains. Malignancy was the only independent predictor (adjusted OR 4.12, 95% CI 1.51–13.10), but it did not remain significant after multiplicity adjustment (BH p = 0.061).

Conclusion: The published Thoracoscore risk-band probabilities should not be used for individual consent at this centre. Recalibration reduces calibration error, but discrimination remains inadequate for reliable individual prediction.

1. Introduction

Thoracic surgery occupies an uncomfortable position in perioperative risk arithmetic. The operations are anatomically major, the physiological reserve of the patients who need them is frequently already eroded by tobacco exposure, malignancy or chronic suppurative lung disease, and the incision itself compromises the very mechanics on which postoperative recovery depends.[1] Even in high-volume European and North American practice, postoperative pulmonary complications follow a substantial minority of non-cardiac thoracic procedures, and complications that arise after apparently successful surgery carry a mortality signal that persists for years.[2–4] Even among candidates whom the American College of Chest Physicians classifies as no worse than moderate risk, morbidity and mortality after lung resection in the European Society of Thoracic Surgeons database are far from negligible.[5] Against that background the surgeon is repeatedly asked, in the clinic and at the bedside, to convert a heterogeneous clinical impression into a single number that a patient and family can weigh.

Risk-prediction models exist to make that conversion defensible. Thoracoscore, derived by Falcoz and colleagues from 15,183 French operations and published in 2007, was the first multivariable model built specifically for general thoracic surgery and remains the most widely deployed one in the United Kingdom and continental Europe.[6,7] It combines nine preoperative and intraoperative descriptors – age band, sex, American Society of Anesthesiologists (ASA) class, performance status, dyspnoea grade, surgical priority, procedure class, diagnostic group and a composite comorbidity score – into a logistic prediction of in-hospital death, and reported a c-index of 0.85 in derivation and 0.86 in internal testing.[6] A subsequent update on a much larger Epithor cohort re-estimated the coefficients and confirmed the general architecture.[8]

External validation, however, has been persistently disappointing. Independent United Kingdom series have returned areas under the curve of 0.68 and 0.69 for open lung resection and as low as 0.44 in a pneumonectomy-only cohort, and several groups have documented systematic overestimation of risk in contemporary practice where minimally invasive access and enhanced recovery pathways have lowered baseline mortality well below the 2002–2005 derivation era.[9,10] An external validation of five published models for cardiopulmonary morbidity in an independent lung-resection population found that none reproduced its derivation performance, and a formal audit of the Eurolung models by their own developers reached the same conclusion.[11,12] Contemporary preoperative-evaluation guidance accordingly treats a published risk figure as provisional until it has been checked locally.[13] The methodological literature has converged on the same point from the other direction: a model is never validated in the

abstract, only in a specified setting, and calibration – not discrimination – is the property that decides whether a predicted probability may be spoken aloud to a patient, and models published without adequate validation are the norm rather than the exception.[14–17]

Almost all of that evidence comes from health systems whose case mix, referral thresholds and critical-care capacity differ profoundly from those of an Indonesian tertiary referral centre. Lung cancer in Indonesia presents late and is managed within a resource envelope that constrains both preoperative optimisation and postoperative escalation.[18,19] International prospective cohorts have repeatedly shown that postoperative mortality after cancer surgery is higher in low- and middle-income settings even after adjustment for patient factors.[20,21] Newer models built specifically to predict postoperative complications in resectable non-small-cell lung cancer share the same transportability problem.[22] If a model derived in France is applied unmodified in Palembang, the arithmetic that reaches the consent conversation may be badly wrong in a direction that matters.

To our knowledge no study has reported the calibration of Thoracoscore in an Indonesian population. The original analysis of the present cohort tested only whether a rank correlation existed between the score band and survival status, and concluded that none did. That question, while reasonable, cannot detect the failure mode that most threatens patients, because a model can be perfectly ordered and still be wrong about the absolute level of risk by an order of magnitude. We therefore re-analysed the cohort at patient level as a formal external validation study, assessing discrimination and calibration separately, auditing the integrity of the score as it was actually recorded, quantifying what the sample could and could not have detected, and deriving a locally recalibrated form for immediate clinical use. The work was conducted jointly by the Departments of Surgery, of Thoracic, Cardiac and Vascular Surgery, and of Pharmacology of the Faculty of Medicine, Universitas Sriwijaya.

2. Methods

2.1. Study design, setting and reporting standard

This was a retrospective observational cohort study reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement and, because the object of study is a prediction model rather than an aetiological association, additionally structured along the TRIPOD+AI recommendations for the reporting of external validation.[23,24] We state at the outset which items of those checklists we could not meet, because a claim of compliance that is not delivered is worse than no claim at all. The source register contains no operative variable, so STROBE items on exposure definition are met only in the weak sense that the exposure is a score; the full model specification required by TRIPOD item 4b is a band table rather than an equation, for the reason given in section 2.3; and the pre-exclusion denominator required by STROBE item 13 could not be recovered from the retrieval log, so the flow diagram in **Figure 1** carries counts only from the point of eligibility assessment onward. The setting was Dr. Mohammad Hoesin General Hospital (RSUP Dr. Mohammad Hoesin), Palembang, the tertiary referral hospital of South Sumatra and the teaching hospital of the Faculty of Medicine, Universitas Sriwijaya. Thoracic operations at this centre are performed by the Division of Thoracic, Cardiac and Vascular Surgery within the Department of Surgery, and the case mix includes both elective oncological resection and urgent operations for suppurative and traumatic disease. The participant flow is summarised in **Figure 1**.

2.2. Participants

All adults aged 18 years or older who underwent a thoracic surgical procedure between 1 January and 31 December 2025 and who were subsequently managed in the general intensive care unit or the surgical ward were eligible. Case ascertainment used ICD-9-CM thoracic procedure codes applied to the hospital medical-record system, and sampling was total rather than fractional: every consecutive eligible record in the 12-month window was retrieved. Records were excluded when death occurred intraoperatively before intensive-care or ward admission, when the patient was discharged against medical advice or transferred to another institution before the index admission concluded (so that the outcome could not be ascertained), or when the record was incomplete or unretrievable for any recorded Thoracoscore component. One hundred and six operations formed the analytic cohort.

Two features of this ascertainment require explicit statement. First, the number of records screened before exclusion was not preserved in the retrieval log, so we cannot demonstrate that the 106 operations exhaust the eligible population, and the description of the cohort as complete should be read as an intention rather than a verified fact. Second, the exclusion of patients discharged against medical advice is not neutral in Indonesian practice, where *pulang paksa* is one route by which a moribund patient leaves hospital. If such episodes were excluded, deaths have been removed from the numerator and the observed mortality reported below is, if anything, an underestimate. The counts were not recoverable and the direction of the bias, though signable, cannot be quantified here.

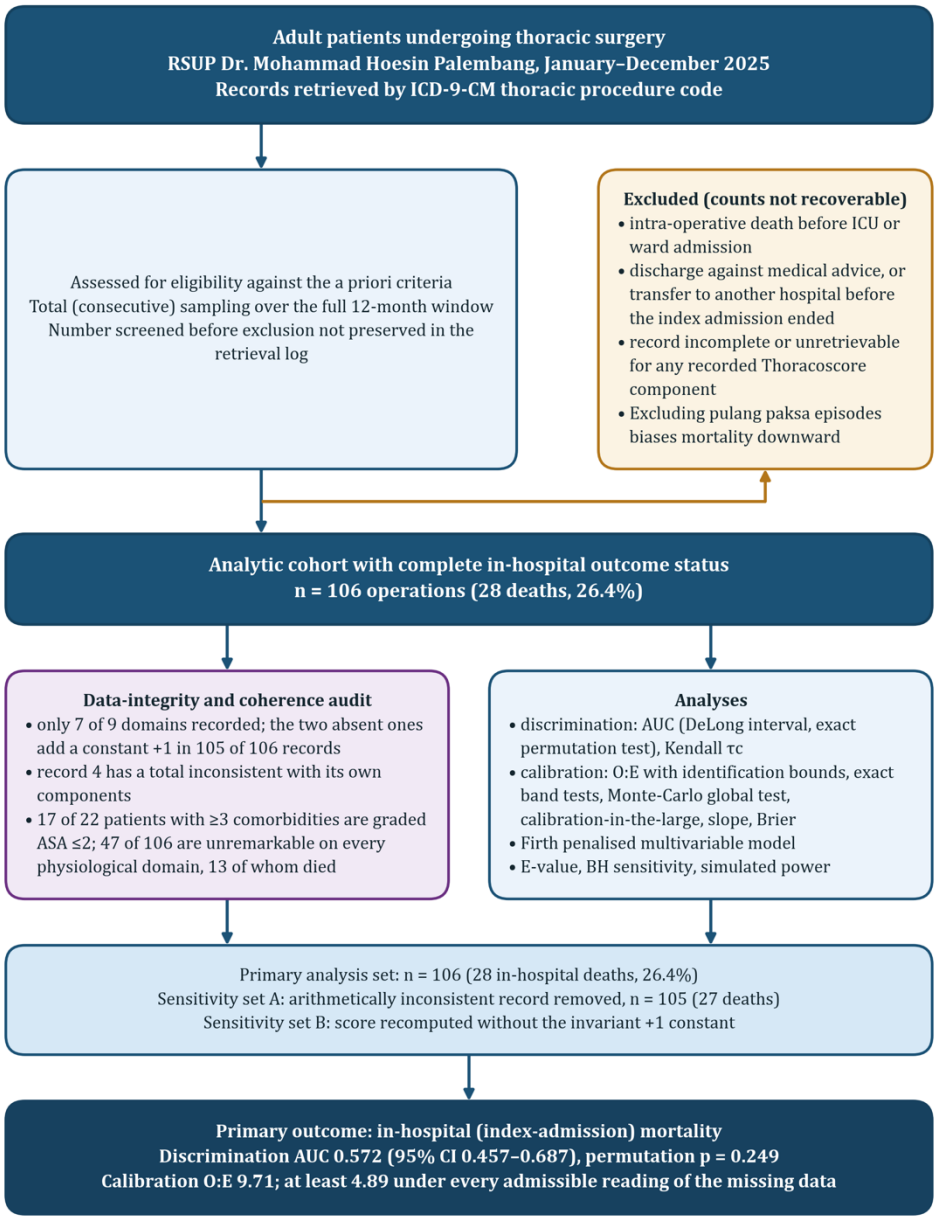

Figure 1. STROBE participant-flow and analysis diagram. All consecutive eligible thoracic operations performed at RSUP Dr. Mohammad Hoesin Palembang between January and December 2025 were retrieved by ICD-9-CM procedure code. The pre-specified data-integrity audit and the three analysis sets are shown.

2.3. Predictor: the Thoracoscore

Thoracoscore assigns points across nine domains and converts the resulting total into a predicted probability of in-hospital death.[6,25] In the present cohort the source register recorded seven of the nine domains at patient level: age band (<55, 55–65, >65 years; 0, 1 and 2 points), female sex (1 point), ASA class ≥3 (1 point), performance status ≥3 (1 point), dyspnoea grade ≥3 (1 point), malignant diagnostic group (1 point) and a comorbidity score (0, ≤2 and ≥3 comorbidities; 0, 1 and 2 points). Surgical priority and procedure class were not recorded at patient level.

Because this omission has consequences for interpretation, we audited it explicitly rather than assuming it away. We reconstructed the seven-component sum for every record and compared it with the total that had been entered in the register. We further pre-specified a duplicate-record check on the complete covariate signature. The results of that audit are reported in the Results and drove two of the three analysis sets.

What was validated, and what was not. Thoracoscore in its original form is a logistic equation returning a patient-level probability.[6] That equation cannot be evaluated here, because two of its nine covariates were never recorded and no reconstruction of them from the register is possible. What is in routine use at this centre, and what we therefore validated, is the five-row points-to-risk band table: 0–2 points, <1%; 3–4 points, 1–3%; 5–6 points, 3–6%; 7–9 points, 6–12%; ≥10 points, >12%.[25] Every calibration statistic in this paper is a property of that band table applied to a seven-domain point total, not of the Falcoz equation, and we have written the title, the abstract and the conclusion to say so. A reader who wants the performance of the original equation in an Indonesian

population will not find it here; the prospective study described in section 4.10 is designed to provide it.

For the primary calibration analysis each patient was assigned the mid-point of the relevant band. Because a band is an interval rather than a point, the expected number of deaths is only partially identified. We therefore computed sharp bounds over two sources of indeterminacy simultaneously. The first is the band interval itself: each patient may be evaluated anywhere between the lower and the upper limit of their band. The second is the two unrecorded domains: surgical priority and procedure class each score either 0 or 1, so a patient's true total lies between the seven-domain sum and that sum plus two, and no other value is admissible. Minimising and then maximising the expected count over both sources jointly gives an interval outside which the true O:E ratio cannot lie, whatever the two missing fields contained.

We record explicitly that the source thesis reported the total point score with a percent sign attached, describing a median Thoracoscore of "3%" with a range of "1%-9%" and risk strata labelled "<2%", "2%-5%" and ">5%-10%". Those figures are point totals, not predicted probabilities; the corresponding predicted probabilities under the scoring table are an order of magnitude smaller. All results below are expressed in points, with the band-derived probability stated separately whenever calibration is at issue.

2.4. Outcome

The outcome was all-cause in-hospital mortality during the index admission, ascertained from the discharge summary and coded as a binary variable. Patients discharged alive were classified as survivors irrespective of subsequent events, since no post-discharge follow-up was available. No outcome data were missing in the analytic cohort by construction of the inclusion criteria.

2.5. Statistical analysis

Analyses were performed in Python 3.11 (NumPy, SciPy, pandas) with a fixed random seed of 20260823 so that every resampling result is exactly reproducible; the analysis script and the de-identified patient-level dataset are available from the corresponding author. Two-sided alpha was 0.05 throughout and p-values are reported to three decimal places.

Descriptive statistics. Categorical variables are given as counts with percentages and, for proportions of clinical interest, with 95% Wilson score intervals, which behave correctly at the small denominators encountered in the extreme risk bands. The Thoracoscore total is summarised by its median and range because it is an integer score on a bounded scale.

Reproduction audit. Before conducting any new analysis we recomputed every statistic reported in the source document. Stuart's tau-c was implemented directly from the concordance and discordance counts of each contingency table and cross-checked against an independent implementation.

Discrimination. The area under the receiver-operating-characteristic curve was estimated as the Mann-Whitney statistic with ties handled by the mid-rank convention and its 95% confidence interval by the DeLong method, which we verified against a 20,000-replicate stratified bootstrap. The DeLong variance is estimated under the alternative and is mildly anti-conservative at these sample sizes; in 60,000 simulated null cohorts drawn from the observed nine-level score distribution its empirical type I error was 5.7% rather than 5%. The test of the AUC against the null value of 0.5 is therefore reported from a 50,000-replicate exact permutation of the outcome labels, with the tie-corrected null-variance normal test as a cross-check. The AUC of the continuous point total was compared with that of the three-band categorisation by the paired DeLong test, which respects the fact that both are computed in the same patients. Component-wise AUCs were computed identically, but they are not comparable with the AUC of the nine-level total and we do not present them as such: for a binary predictor the area under the curve is the mean of sensitivity and specificity, and its attainable maximum is bounded by the prevalence of exposure. At the prevalences observed here that ceiling is 0.661 for ASA class ≥3, 0.732 for performance status ≥3 and 0.750 for dyspnoea grade ≥3, and each component AUC should be read against its own ceiling rather than against unity.

Calibration. Calibration was assessed at four resolutions. The observed-to-expected ratio across the whole cohort is reported with an exact Poisson interval for the observed count and, separately, with the sharp identification interval described in section 2.3; the Poisson interval conditions on the expected count being known, which the identification interval shows it is not, so the two must be read together rather than either alone. Band-wise calibration is the observed proportion with its Wilson interval against the band-specific predicted probability, tested within each band by an exact one-sided binomial test. A global goodness-of-fit test was performed by Monte-Carlo simulation of 400,000 cohorts under the model rather than by a Hosmer-Lemeshow statistic; with expected counts below one in every band the chi-square approximation does not hold, and for an externally specified model with pre-specified groups the conventional degrees-of-freedom correction does not apply either. Calibration-in-the-large is the intercept of a logistic model with the linear predictor entered as an offset, that is with the slope fixed at unity, which is the quantity that measures systematic underprediction. The calibration slope is the coefficient when the linear predictor is entered freely, estimated by ordinary maximum likelihood, with a profile-likelihood interval and profile likelihood-ratio tests against both unity and zero. A slope below unity means the predicted risks are too widely spread on the logit scale relative to the observed spread; it is a separate matter from the level of risk, and we report the logit-scale spreads directly so that the two are not confused. Overall accuracy is summarised by the Brier score and by the scaled Brier score relative to the cohort event rate, a negative value of which means that the model is less accurate than simply quoting the base rate.[15,16]

Association analyses. Two-by-two associations were tested by Fisher's exact test with conditional maximum-likelihood odds ratios and exact confidence intervals; ordered exposures were tested by the Cochran-Armitage trend statistic, which preserves the ordering that an omnibus contingency test discards. Odds ratios are the conditional maximum-likelihood estimates that match the exact

conditional intervals reported beside them. Because several tests were performed in a single family, p-values were additionally adjusted by the Benjamini-Hochberg procedure. The definition of that family is a judgement rather than a fact, and the adjusted value depends on it, so we report the adjustment under three defensible definitions: all nine tests, the seven tests of the individual recorded domains, and the five two-by-two tests alone.

Multivariable modelling. With 28 events and seven candidate predictors the events-per-variable ratio is 4.0, well below the conventional threshold of ten, and ordinary maximum likelihood would be expected to produce inflated coefficients and unstable standard errors.[26] We therefore fitted a Firth penalised logistic model, reporting adjusted odds ratios with 95% profile penalised-likelihood confidence intervals and penalised likelihood-ratio tests rather than Wald statistics. The apparent AUC of the refitted model was corrected for optimism by 2,000 bootstrap replications of the entire fitting procedure, and, because the optimism bootstrap is known to be insufficiently pessimistic for a shrunken model in a small sample, additionally by 50 repetitions of 10-fold cross-validation and by out-of-bag bootstrap; all three are reported.[16,26]

Clinical utility. Decision-curve analysis compared the net benefit of the published score, of a locally recalibrated version of it and of the default strategies of treating all and treating no patients as high risk, on a grid of risk thresholds from 2% to 45% in steps of one percentage point.[27] Because the recalibration was fitted in the same 106 patients, the recalibrated curve was corrected for optimism by 300 bootstrap replications of the whole recalibrate-and-evaluate procedure, and only the corrected curve is interpreted. Threshold performance for every possible integer cut-point is tabulated, and the apparent Youden index of the best cut-point was corrected for the optimism induced by selecting that cut-point in the same data, again by 2,000 bootstrap replications.

Sensitivity of a null result. Rather than assert that a null result reflects inadequate power, we quantified what the design could detect. The power of the AUC test was obtained by simulating 40,000 cohorts per scenario in which case scores were drawn from a reweighting of the *observed* nine-level score distribution rather than from a binormal model, so that the heavy ties of an integer score are represented; power is reported both at the nominal critical value and at the size-corrected critical value implied by the empirical type I error above, and, because power at a given AUC depends on how the discriminating signal is distributed across the score range, under three different shapes of alternative. The minimum detectable odds ratio was obtained by direct simulation of Fisher's exact test at the observed exposure prevalences, not by a normal approximation, which is unreliable at these cell counts. The robustness of the one positive association to unmeasured confounding was expressed as an E-value computed from the *adjusted* estimate and, as the cited best-practice guidance requires, also at the confidence limit closest to the null; a fragility index is reported for completeness with its exact definition stated.[28]

Analysis sets. The primary analysis set comprised all 106 records. Sensitivity set A excluded the one record whose recorded total is inconsistent with its own components. Sensitivity set B recomputed the score without the invariant constant identified in the data-integrity audit. All three are reported.

2.6. Refusals stated in advance

Three analyses that a reader might reasonably expect are deliberately absent, and we state the reason rather than allow their absence to be read as oversight. First, no time-to-event analysis is presented: the register records vital status at discharge only, with neither date of death nor any post-discharge follow-up, so a Kaplan-Meier estimate would require dates that do not exist. Second, no Clavien-Dindo distribution is presented, because postoperative complications were not graded in the source register; the grading system is recommended for the prospective phase of this work and appears in the algorithm of Figure 6 for that reason.[29,30] Third, and most importantly for the pharmacological component of this collaboration, no drug-exposure, dose, serum-concentration or biomarker analysis is presented, because no such variable was recorded. The pharmacological contribution to this paper is therefore explicitly methodological and bias-analytic rather than empirical, and is described as such below.

2.7. Pharmacological contribution

The Department of Pharmacology contributed to this study in three specific and verifiable ways. First, it framed the class of unmeasured perioperative determinants against which the observed associations must be judged: transfusion exposure and perioperative anaemia, intraoperative ventilation and haemodynamic strategy, analgesic regimen, antimicrobial choice and the handling of chronic antithrombotic therapy are all established modifiers of outcome after thoracic surgery and none is captured by any Thoracoscore component.[31–35] Second, it specified the quantitative bias analysis by which the single positive association in this cohort is tested against that unmeasured class, using the E-value framework to state how strong an unmeasured perioperative exposure would have to be, on the risk-ratio scale and in association with both the indication and death, to explain the finding away.[28] Third, it specified the pharmacological data model for the prospective successor study, in which medication reconciliation, antithrombotic bridging, analgesic technique, vasopressor and fluid exposure and antimicrobial de-escalation are recorded as candidate predictors alongside the Thoracoscore components, and cardiovascular assessment follows current European guidance.[36–38]

2.8. Ethics

This study received ethical approval from the Health Research Ethics Committee of Dr. Mohammad Hoesin Central General Hospital, Palembang, Indonesia (Ref. No. DP.04.03/D.XVII.9.4/ETIK/166/2026), granted after the close of the accrual window as is customary for a retrospective review of existing records. Because the study analysed medical records already created in the course of clinical care, from which identifiers were removed at the point of extraction, and because 28 of the 106 patients died during the index admission, the committee waived the requirement for

individual written informed consent; no participant was approached and no additional clinical data were generated. Permission to access the medical records was granted by the Education and Research Directorate of the same institution. The study was conducted in accordance with the Declaration of Helsinki. Patient identifiers were replaced by initials at the point of extraction and no identifying information appears in this report or in the dataset shared with reviewers.

3. Results

3.1. Cohort characteristics

One hundred and six thoracic operations met the inclusion criteria. Seventy-nine patients (74.5%) were male and 27 (25.5%) female. Half the cohort (53 patients, 50.0%) fell in the 55–65 year band, 35 (33.0%) were younger than 55 years and 18 (17.0%) were older than 65 years. Sixty-four operations (60.4%) were performed for a malignant indication and 42 (39.6%) for benign disease. The cohort was, by the metrics the score uses, a fit one: 97 patients (91.5%) were ASA class 2 or below, 93 (87.7%) had a performance status of 2 or better, 92 (86.8%) had a dyspnoea grade of 2 or less, and 54 (50.9%) carried no recorded comorbidity at all. Full characteristics with band-specific mortality are given in **Table 1**.

Table 1. Demographic, clinical and Thoracoscore-component characteristics of the cohort, with band-specific in-hospital mortality (N = 106)

Characteristic	n (%)	Deaths, n	Mortality % (95% CI)
Sex			
Male	79 (74.5)	19	24.1 (16.0–34.5)
Female	27 (25.5)	9	33.3 (18.6–52.2)
Age band (years)			
<55	35 (33.0)	8	22.9 (12.1–39.0)
55–65	53 (50.0)	15	28.3 (18.0–41.6)
>65	18 (17.0)	5	27.8 (12.5–50.9)
Diagnostic group			
Benign	42 (39.6)	5	11.9 (5.2–25.0)
Malignant	64 (60.4)	23	35.9 (25.3–48.2)
ASA physical status			
≤2	97 (91.5)	26	26.8 (19.0–36.4)
≥3	9 (8.5)	2	22.2 (6.3–54.7)
Performance status (ECOG)			
≤2	93 (87.7)	24	25.8 (18.0–35.5)
≥3	13 (12.3)	4	30.8 (12.7–57.6)
Dyspnoea grade (MRC)			
≤2	92 (86.8)	23	25.0 (17.3–34.7)
≥3	14 (13.2)	5	35.7 (16.3–61.2)
Comorbidity score			
0 comorbidities	54 (50.9)	14	25.9 (16.1–38.9)
≤2	30 (28.3)	9	30.0 (16.7–47.9)
≥3	22 (20.8)	5	22.7 (10.1–43.4)
Thoracoscore risk band (as originally assigned)			
Low (1 point)	7 (6.6)	0	0.0 (0.0–35.4)
Intermediate (2–5 points)	80 (75.5)	24	30.0 (21.1–40.8)
High (6–9 points)	19 (17.9)	4	21.1 (8.5–43.3)
Thoracoscore total, points	median 3 (range 1–9)		
All patients	106 (100.0)	28	26.4 (19.0–35.5)

Notes: ASA, American Society of Anesthesiologists; ECOG, Eastern Cooperative Oncology Group; MRC, Medical Research Council. Percentages are column percentages of the whole cohort; mortality percentages are row percentages within the stratum, with 95% Wilson score intervals. The comorbidity score counts smoking, prior malignancy, chronic obstructive pulmonary disease, arterial hypertension, cardiac disease, diabetes mellitus, peripheral vascular disease, obesity and alcohol use. No operative variable – procedure, urgency, approach, extent of resection, duration or blood loss – was recorded in the source register and none can therefore be tabulated.

The Thoracoscore total had a median of 3 points and ranged from 1 to 9. Under the three-band scheme used in the source analysis, 7 patients (6.6%) were classified low risk, 80 (75.5%) intermediate risk and 19 (17.9%) high risk. Twenty-eight patients died before discharge, an in-hospital mortality of 26.4% (95% CI 19.0–35.5).

3.2. Data-integrity audit

The audit specified in the Methods produced two findings that condition everything that follows, and both are exactly reproducible from the appended dataset (**Table 2**).

Table 2. Data-integrity and predictor-coherence audit of the source register.

Audit item	Finding	Consequence for the analysis
Thoracoscore domains recorded at patient level	7 of 9; surgical priority and procedure class absent	The published logistic equation cannot be evaluated; the five-row risk-band table was validated instead
Recorded total minus the sum of the 7 recorded components	+1 point in 105 of 106 records; +0 in 1 record (record no. 4)	The two absent domains contributed a constant, not a variable
Effect of the constant on rank statistics	AUC 0.572 as recorded vs 0.579 with the constant removed	None in principle; the small observed difference arises only because one record does not carry the constant
Effect of the constant on risk-band membership	33 of 106 patients (31.1%) change band when it is removed	A third of the cohort carries a risk label that depends on two unrecorded fields
Arithmetically inconsistent record	Record 4 is the only record whose total is not the component sum plus the constant	Analysed as sensitivity set A with the record removed; no calibration or discrimination result changed
Is that record a duplicate of record 50?	It matches on initials, outcome and all 7 domains, but 83 of 106 records share a 7-domain signature and such a match arises by chance in 49% of 10,000 label permutations	The duplication claim is not supported; the transcription inconsistency is
ASA class versus recorded comorbidity count	17 of 22 patients with ≥3 comorbidities are graded ASA ≤2; 4 patients graded ASA ≥3 have no recorded comorbidity	The physiological fields are internally incoherent; systematic under-grading is a competing explanation for the miscalibration
Patients unremarkable on every recorded physiological domain	47 of 106 (44.3%) are ASA ≤2, ECOG ≤2, dyspnoea ≤2 and comorbidity-free; 13 of them died (27.7%)	Mortality in the apparently fittest patients is indistinguishable from the cohort as a whole
Malignant indication versus ASA class	only 6 of 64 patients operated for malignancy are graded ASA ≥3	Consistent with the same under-grading

Notes: ASA, American Society of Anesthesiologists; AUC, area under the receiver-operating-characteristic curve; ECOG, Eastern Cooperative Oncology Group. Every item is exactly reproducible from the patient-level dataset.

First, the recorded Thoracoscore total equals the sum of the seven recorded components plus exactly one point in 105 of 106 records, and equals that sum with no addition in the single remaining record. The two unrecorded domains – surgical priority and procedure class – therefore contributed a constant, not a variable, across essentially the whole cohort. Two consequences follow. Because an additive constant is a monotone transformation it cannot alter a rank-based statistic, so the AUC would be unchanged were the constant removed from every record; the small observed difference (0.572 with the constant versus 0.579 without) arises entirely because one record does not carry it, and that record's rank therefore moves. Because band boundaries are absolute, by contrast, the same constant moves patients across them: recomputing the bands without it reclassifies 33 of 106 patients (31.1%), converting a cohort of 7 low-risk, 80 intermediate-risk and 19 high-risk patients into one of 32, 63 and 11 respectively. A risk label that a third of patients would not have received had two fields been completed is not a label that should be quoted at consent.

Second, record 4 is the one record whose total is inconsistent with its own components, and it also matches record 50 on initials, outcome and all seven recorded domains. We initially read that match as evidence of duplication and, on the reviewers' prompting, tested it. It is not evidence. The seven domains define only 288 possible signatures and the observed distribution is highly concentrated: 83 of 106 records share their seven-domain signature with at least one other record and 75 share signature and outcome. Permuting the initials strings across records 10,000 times, at least one record matches another on initials, outcome and all seven domains in 49.3% of permutations. A coincidence that occurs by chance half the time is not a duplicate. What distinguishes record 4 is the arithmetic inconsistency of its total, which is a transcription finding; all analyses were repeated with it removed (sensitivity set A) and nothing changed.

Third, and not part of our original audit, the recorded predictor values are internally incoherent in a way that bears directly on the interpretation of everything that follows (**Table 2**). Seventeen of the 22 patients carrying three or more comorbidities are simultaneously graded ASA 2 or below, and four patients graded ASA 3 or above carry no recorded comorbidity at all. Only 6 of the 64 patients operated for malignancy are graded ASA 3 or above. Most strikingly, 47 patients – 44.3% of the cohort – are recorded as entirely unremarkable on every clinical domain the score uses (ASA ≤2, performance status ≤2, dyspnoea grade ≤2 and no comorbidity), and 13 of those 47 died in hospital, a mortality of 27.7% that is indistinguishable from the cohort as a whole. A patient carrying three or more of the listed comorbidities is ASA 3 by the definition of the class. That 77% of such patients are graded ASA 2 or below is not a case-mix observation; it is evidence that the physiological fields were systematically under-graded or were abstracted from a pre-admission note rather than the anaesthetic record. We return to the consequences in section 4.3.

3.3. Reproduction of the originally reported analysis

Every one of the eight Kendall tau-c coefficients reported in the source document reproduces exactly to three decimal places from the appended patient-level data (**Table 3**). The recomputed p-values differ from the published values by at most 0.037, a discrepancy fully explained by the difference between the asymptotic standard-error formula used by commercial statistical software and the one implemented here; no discrepancy alters any conclusion, and the one significant association (diagnostic group) remains significant on recomputation ($p = 0.006$ versus a published 0.002). The original arithmetic is therefore sound. What the original analysis did not do was test the question that matters clinically.

Table 3. Reproduction of the eight Kendall tau-c analyses reported in the source document.

Variable	Reported τc	Recomputed τc	Agreement	Reported p	Recomputed p	|Δp|
Sex	0.070	0.070	exact	0.368	0.347	0.021
Age band	0.045	0.045	exact	0.611	0.618	0.007
Diagnostic group	0.230	0.230	exact	0.002	0.006	0.004
ASA class	-0.014	-0.014	exact	0.754	0.767	0.013
Performance status	0.021	0.021	exact	0.716	0.705	0.011
Dyspnoea grade	0.049	0.049	exact	0.436	0.399	0.037
Comorbidity score	-0.007	-0.007	exact	0.936	0.938	0.002
Thoracoscore risk band	0.021	0.021	exact	0.738	0.775	0.037
Thoracoscore, underlying integer points (not reported in the source)	-	0.112	new	-	0.249	-

Notes: τc, Stuart's tau-c. All eight coefficients reproduce exactly to three decimal places. The residual differences in p-value reflect the asymptotic standard-error formula used by the original software and alter no conclusion. The final row shows the same statistic computed on the underlying nine-level integer score rather than on the three-band categorisation.

One consequence of the analytical choice made in the source document is worth isolating. The reported association between the Thoracoscore and mortality (tau-c 0.021, p = 0.775 on recomputation) was computed on the three-band categorisation. Computed on the underlying integer score, the same statistic is 0.112 (p = 0.249) – more than five times larger. Categorising a nine-level ordered score into three bands discarded roughly four-fifths of what little association was present. The corresponding loss of discrimination is 0.059 AUC units (0.572 versus 0.514; DeLong z = 1.438, p = 0.150), a difference that is not statistically significant in this sample but is entirely a self-inflicted analytic loss rather than a property of the score.

A third observation belongs here because it concerns our own arithmetic rather than the source document's. Our first draft reported the discrimination p-value from the DeLong variance, which is estimated under the alternative. In 60,000 simulated null cohorts drawn from the observed score distribution that test rejects at 5.7% rather than 5%, and the resulting p-value is correspondingly too small. The correct value, from a 50,000-replicate exact permutation of the outcome labels, is 0.249, and the tie-corrected null-variance normal test agrees to three decimal places. We report 0.249 throughout, note that the discrepancy of 0.031 is of the same order as the discrepancies we identified in the source document, and observe that the DeLong *interval* is unaffected: a 20,000-replicate stratified bootstrap reproduces 0.457–0.686 against the DeLong 0.457–0.687.

3.4. Discrimination

The Thoracoscore total separated survivors from non-survivors no better than chance. The AUC was 0.572 (95% CI 0.457–0.687; permutation p = 0.249 against the null value of 0.5), and the three-band form reached only 0.514 (95% CI 0.432–0.595; p = 0.775). Component-wise, only the diagnostic group carried any discriminatory signal at all (AUC 0.648, 95% CI 0.557–0.739); ASA class (0.491), the comorbidity score (0.495), performance status (0.514), the age band (0.529), dyspnoea grade (0.532) and sex (0.545) were each indistinguishable from a coin toss – and, for the three binary physiological domains, each also far below the ceiling their own exposure prevalence permits (0.661, 0.732 and 0.750 respectively), so the null result is not an artefact of rarity. Receiver-operating characteristic curves are shown in **Figure 2A** and the numerical results in **Table 4**.

Table 4. Discrimination of the Thoracoscore, its components and a locally refitted model for in-hospital mortality.

Score or component	AUC	95% CI	p vs 0.5
Thoracoscore, integer points	0.572	0.457–0.687	0.249 (permutation)
Thoracoscore, three-band form	0.514	0.432–0.595	0.775
Difference (paired DeLong)	0.059	z = 1.438	0.150
Age band	0.529	0.416–0.643	-
Female sex	0.545	0.445–0.645	-
ASA≥3	0.491	0.433–0.549	-
PS≥3	0.514	0.439–0.589	-
Dyspnoea≥3	0.532	0.451–0.612	-
Malignant	0.648	0.557–0.739	-
Comorbidity	0.495	0.382–0.609	-
Locally refitted 7-component model, apparent	0.729	0.617–0.841	-
Bootstrap optimism-corrected	0.646	optimism 0.083	-
50 x 10-fold cross-validated	0.555	SD 0.025	-
Out-of-bag bootstrap	0.578	400 replications	-
Derivation cohort, for reference	0.850	reported in derivation	-

Notes: AUC, area under the receiver-operating-characteristic curve, estimated as the Mann-Whitney statistic with mid-rank tie handling; confidence intervals by the DeLong method, verified against a 20,000-replicate stratified bootstrap. The test against 0.5 is a 50,000-replicate exact permutation of the outcome labels, because the DeLong variance is estimated under the alternative and rejects at 5.7% rather than 5% in 60,000 simulated null cohorts. Three internal-validation routes for the refitted model are reported because they disagree; the cross-validated value is the one we regard as honest. The derivation c-index is quoted from the original publication and was not recomputed.

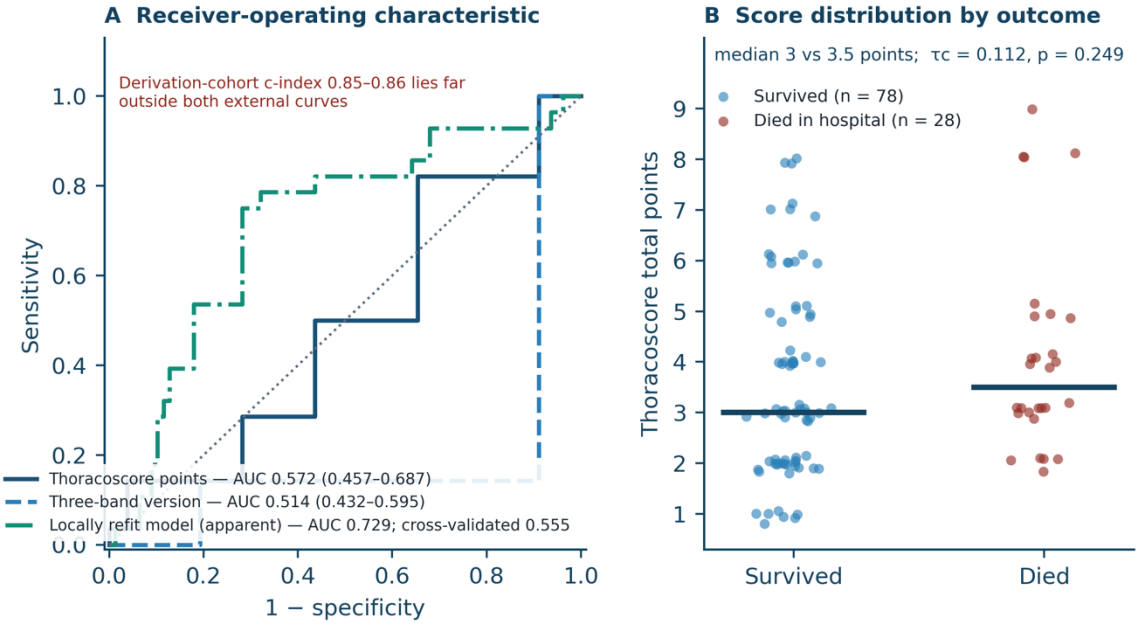

Figure 2. Discrimination of the Thoracoscore for in-hospital mortality. (A) Receiver-operating-characteristic curves for the integer point total, its three-band categorisation and a model refitted locally on all seven recorded components; the upper confidence limit of every curve excludes the c-index of 0.85–0.86 reported in the derivation cohort. (B) Distribution of the point total by outcome with horizontal bars at the group medians; the distributions overlap almost completely.

The upper confidence limit of 0.687 for the full score is the most informative single number in this section. It excludes not only the c-index of 0.85–0.86 reported in the derivation cohort but also the value of 0.70 that is conventionally taken as the threshold of useful discrimination.[6,16] A locally refitted model using all seven recorded components – that is, the best that these variables can do when their weights are estimated in this very cohort – achieved an apparent AUC of 0.729 (95% CI 0.617–0.841). Three internal validation routes disagree about how much of that is real, and we report all three rather than the most flattering. Bootstrap optimism correction gives 0.646; twenty repetitions of 10-fold cross-validation give 0.555 (SD 0.025); out-of-bag bootstrap gives 0.578. The optimism bootstrap is known to be insufficiently pessimistic for a shrunken model in a small sample, and the cross-validated figure is the one we regard as the honest estimate of out-of-sample performance. It is, if anything, slightly below the published score's own 0.572. The limitation is therefore not the French coefficients but the predictors themselves: in this population the recorded Thoracoscore domains do not contain the information that separates the patients who die from those who do not.

Figure 2B shows why. The score distributions of survivors and non-survivors overlap almost completely; the medians differ by half a point (3 versus 3.5) and both groups span essentially the same range.

3.5. Mortality at each level of the score

The three-band collapse conceals a pattern that a surgical reader should see (**Table 5**). Mortality rises from 0 of 7 at 1 point to 5 of 25 (20.0%) at 2 points, 9 of 26 (34.6%) at 3, 6 of 18 (33.3%) at 4 and 4 of 11 (36.4%) at 5. It then falls to 0 of 8 at 6 points and 0 of 4 at 7 points, before rising to 3 of 6 and 1 of 1 at 8 and 9 points. Twelve consecutive patients scoring 6 or 7 points all survived while 4 of the 7 patients scoring 8 or 9 died, a contrast that is itself significant on exact testing (p = 0.009).

Table 5. In-hospital mortality at each level of the Thoracoscore total

Thoracoscore total (points)	n	Deaths	Mortality % (95% CI)	Band predicted %
1	7	0	0.0 (0.0–35.4)	0.5
2	25	5	20.0 (8.9–39.1)	0.5
3	26	9	34.6 (19.4–53.8)	2.0
4	18	6	33.3 (16.3–56.3)	2.0
5	11	4	36.4 (15.2–64.6)	4.5
6	8	0	0.0 (0.0–32.4)	4.5
7	4	0	0.0 (0.0–49.0)	9.0
8	6	3	50.0 (18.8–81.2)	9.0
9	1	1	100.0 (20.7–100.0)	9.0
6–7 points vs 8–9 points	12 vs 7	0 vs 4	Fisher exact p = 0.009	

Notes: 95% Wilson score intervals. Observed mortality is not monotone in the score: twelve consecutive patients scoring 6 or 7 points all survived while 4 of the 7 patients scoring 8 or 9 points died. We interpret this as a signal about the register rather than about the patients; see section 4.3.

We do not regard this as a biological finding. A discontinuity of that size inside a nine-point additive score, in

a cohort whose overall mortality is five to ten times that of any published thoracic series, in a register that we have already shown to omit two fields entirely, to mis-transcribe at least one total and to grade ASA incoherently, is at least as consistent with measurement error as with anything about the patients. It is reported because it is in the data and because a reader is entitled to see it, and it is discussed as an alternative explanation in section 4.3 rather than as a result.

3.6. Calibration

Calibration failed on a scale that discrimination statistics alone would never have revealed (**Table 6**, **Figure 3**). Applying the band mid-points of the scoring table to the observed score distribution yields 2.88 expected deaths. Twenty-eight were observed. The observed-to-expected ratio is 9.71, with an exact Poisson 95% interval on the observed count of 6.45 to 14.03. That interval conditions on the expected count being known, which it is not, so the primary statement of uncertainty is the identification bound rather than the Poisson interval.

Table 6. Calibration of the Thoracoscore risk-band table against observed in-hospital mortality

Thoracoscore band (points)	n	Predicted %	Observed deaths	Observed % (95% CI)	O:E	Exact p
1–2 points	32	0.5	5	15.6 (6.9–31.8)	31.2	<0.001
3–4 points	44	2.0	15	34.1 (21.9–48.9)	17.0	<0.001
5–6 points	19	4.5	4	21.1 (8.5–43.3)	4.7	0.009
7–9 points	11	9.0	4	36.4 (15.2–64.6)	4.0	0.013
Whole cohort	106	2.72	28	26.4 (19.0–35.5)	**9.71**	
Expected deaths at band mid-points			2.88	Poisson 95% CI on O:E 6.45–14.03		
Sharp bound over bands and both unrecorded domains			1.08–5.73	O:E 4.89–25.93		
Sharp bound over the bands alone			1.67–4.1	O:E 6.83–16.77		
Global Monte-Carlo test (400,000 cohorts under the model)			max simulated deaths 14	observed 28		<0.001
Calibration-in-the-large (slope fixed at 1)			2.766	SE 0.238; odds x15.9		
Calibration slope (maximum likelihood)			0.291	95% profile CI -0.152 to 0.754		0.003 vs 1
Slope tested against zero			$\chi^2 = 1.65$			0.199
Logit-scale spread, predicted vs observed			2.98 vs 1.13	ratio 0.38 ≈ the fitted slope		
Brier / null Brier / scaled Brier			0.2492 / 0.1944	-0.282		

Notes: O:E, observed-to-expected ratio. Predicted mortality is the mid-point of the band specified in the scoring table in local use (0–2 points, <1%; 3–4, 1–3%; 5–6, 3–6%; 7–9, 6–12%). Exact p-values are one-sided binomial tests within band. No Hosmer-Lemeshow statistic is reported: the expected counts are 0.16, 0.88, 0.86 and 0.99, far below the value at which the chi-square approximation holds, and the conventional degrees-of-freedom correction does not apply to an externally specified model with pre-specified groups. A negative scaled Brier score indicates that the model is less accurate than the cohort event rate alone.

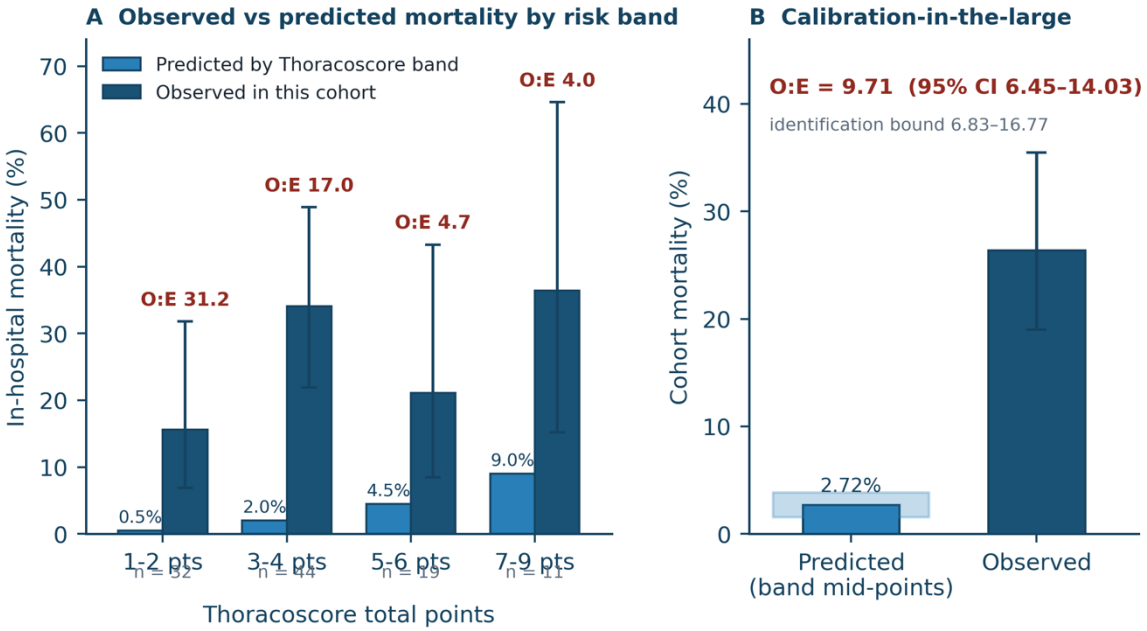

Figure 3. Calibration of the Thoracoscore in this cohort. (A) Observed in-hospital mortality with 95% Wilson intervals against the mortality predicted by the risk band, with the band-specific observed-to-expected (O:E) ratio above each pair. Observed mortality is not monotone across bands and the underprediction is worst in the lowest band. (B) Calibration-in-the-large; the shaded region is the sharp identification bound on the predicted rate implied by evaluating every patient at the limits of their band.

Minimising and then maximising the expected count over both sources of indeterminacy jointly – the width of each risk band, and the two unrecorded domains, which can add 0, 1 or 2 points to any patient's true total and no more – gives an expected count between 1.08 and 5.73 and hence an O:E ratio between 4.89 and 25.93. This is the honest floor: whatever surgical priority and procedure class contained, and wherever inside its band each patient is placed, the band table underpredicts by at least a factor of 4.9. Under the more restrictive assumption that the recorded seven-domain total is itself the true total, the bound narrows to 6.83–16.77. Shifting every recorded total by one point in either direction moves the point estimate from 9.71 to 13.73 (downward shift) or 6.76 (upward shift), so the finding is not delicately dependent on the constant identified in the audit.

A global goodness-of-fit test was performed by simulation rather than by a Hosmer-Lemeshow statistic, which is not usable here: the expected counts in the four occupied bands are 0.16, 0.88, 0.86 and 0.99, all far below the value at which the chi-square approximation holds, and for an externally specified model with pre-specified groups the conventional degrees-of-freedom correction does not apply. In 400,000 cohorts simulated under the model the mean number of deaths was 2.89 and the largest number ever generated was 14, against 28 observed ($p < 0.00001$). That single sentence carries the finding more securely than any chi-square statistic could.

The failure is present in every occupied band and is worst where the score is most reassuring. Among the 32 patients scoring 1–2 points, for whom the table predicts a mortality below 1%, five died (15.6%, 95% CI 6.9–31.8), an O:E of 31.2. Among the 44 patients scoring 3–4 points (predicted 1–3%), fifteen died (34.1%, 95% CI 21.9–48.9), an O:E of 17.0. The 19 patients scoring 5–6 points had an observed mortality of 21.1% against a predicted 4.5% (O:E 4.7) and the 11 scoring 7–9 points 36.4% against a predicted 9.0% (O:E 4.0). Exact one-sided binomial tests within each band reject the band-specific predicted probability in all four: $p < 0.000001$, $p < 0.000001$, $p = 0.009$ and $p = 0.013$ respectively.

Two further features of this pattern deserve emphasis. First, the observed mortality is not monotone across the bands: it rises from 15.6% to 34.1%, falls to 21.1% and rises again to 36.4%. The score does not merely misplace the level of risk, it partially misorders it. Second, the O:E ratio falls as the score rises, from 31.2 in the lowest band to 4.0 in the highest. The underprediction is worst precisely where the score is used to reassure, and the reassurance offered to a patient scoring 1 or 2 points at this centre – a predicted risk below one per cent against an observed risk of roughly one in six – is the single most clinically dangerous finding of this study.

Formal recalibration statistics separate two distinct failures that the O:E ratio alone conflates. Calibration-in-the-large – the intercept of a model in which the linear predictor is entered as an offset, so that the slope is fixed at unity – is 2.766 (SE 0.238), a 15.9-fold shift in the odds of death relative to what the band table predicts. That is the magnitude of the systematic underprediction, and it is the single most important number in this paper. The calibration slope, estimated freely by maximum likelihood, is 0.291 with a 95% profile-likelihood interval of -0.152 to 0.754. It differs from unity (profile likelihood-ratio chi-square 8.66, $p = 0.003$) and does not differ from zero (chi-square 1.65, $p = 0.199$).

The direction of that slope requires care, because it is easy to state backwards. A slope below unity means the predicted risks are too widely spread on the logit scale relative to the observed spread, not too narrowly. Here the predicted linear predictor spans 2.98 logit units while the observed band-level logits span 1.13, a ratio of 0.38 that is essentially the fitted slope. The band table is therefore simultaneously too confident about how much its bands differ from one another and far too low about all of them. That the slope cannot be distinguished from zero is a second, independent statement of the same finding reported in section 3.4: the linear predictor carries no demonstrable information about who dies.

Two recalibrated forms of the band table were fitted for reference and are reported in **Table 7** alongside the observed rates and the discredited multiplicative rule discussed in section 4.8; neither reproduces the observed pattern, for the reason given there.

Table 7. Published, multiplicatively rescaled and formally recalibrated band probabilities against observed mortality

Thoracoscore band (points)	Published %	x local O:E of 9.71 %	Intercept-only recalibration %	Intercept + slope %	Observed %
1–2 points	0.5	4.9	7.4	19.5	15.6
3–4 points	2.0	19.4	24.5	26.7	34.1
5–6 points	4.5	43.7	42.8	31.7	21.1
7–9 points	9.0	87.4	61.1	36.5	36.4

Notes: The multiplicative rule – multiplying the published band probability by the cohort observed-to-expected ratio – was proposed in an earlier version of this manuscript and is withdrawn: it still understates risk threefold in the lowest band and returns 87.4% in the highest band against an observed 36.4%. The two fitted recalibrations correct the level of risk but cannot reproduce a non-monotone observed pattern with a monotone transformation of a monotone predictor, and the two-parameter version is very nearly a constant close to the cohort base rate.

The Brier score of the published model was 0.2492 against 0.1944 for the trivial strategy of assigning every patient the cohort event rate, giving a scaled Brier score of -0.282. A negative scaled Brier score is unambiguous: in this cohort the band table is a less accurate probability estimate than saying nothing more than "about one in four".

3.7. Predictors of in-hospital mortality

On univariable exact testing, a malignant indication was the only variable associated with death: 23 of 64 patients operated for malignancy died (35.9%) against 5 of 42 operated for benign disease (11.9%), giving a conditional maximum-likelihood odds ratio of 4.10 (95% exact CI 1.34–15.24; $p = 0.007$), a risk ratio of 3.02 and an absolute risk

difference of 24.0 percentage points. We do not express that difference as a number needed to harm, because a malignant indication for surgery is a patient characteristic and not an intervention that can be withheld. Sex, ASA class, performance status, dyspnoea grade, the age band and the comorbidity score were all null, and the Cochran-Armitage trend statistic across the ordered Thoracoscore bands was likewise null (p = 0.704). The Benjamini-Hochberg-adjusted p-value for the malignancy association depends on how the test family is defined, and no definition is privileged: across all nine tests it is 0.061, across the seven tests of the individual recorded domains 0.048, and across the five two-by-two tests alone 0.034 (**Table 8**). Two of the nine tests are tests on the score itself, which is a deterministic function of the seven components already being tested, so the nine-test family is internally redundant; but we did not pre-specify a family and we therefore report the sensitivity rather than choosing the definition that suits the result. The honest summary is that the association sits on the conventional threshold and moves across it according to a defensible analytic choice.

Table 8. Univariable and Firth penalised multivariable analysis of the seven recorded Thoracoscore components.

Variable	Deaths / n exposed	Univariable estimate (95% CI)	p	p (BH)	Adjusted OR (95% CI)	p
Female sex	9 / 27	OR 1.58 (0.53–4.46)	0.448	0.953	1.80 (0.65–4.93)	0.110
Malignant diagnosis	23 / 64	OR 4.1 (1.34–15.24)	0.007	0.061	4.12 (1.51–13.10)	0.002
ASA ≥3	2 / 9	OR 0.78 (0.07–4.47)	1.000	1.000	-	-
Performance status ≥3	4 / 13	OR 1.28 (0.26–5.11)	0.741	0.953	-	-
Dyspnoea score ≥3	5 / 14	OR 1.67 (0.39–6.21)	0.515	0.953	-	-
Age band (trend)	-	trend z = 0.477	0.633	0.953	0.90 (0.41–1.96)	0.172
Comorbidity score (trend)	-	trend z = -0.152	0.879	0.989	0.76 (0.39–1.43)	0.086
Thoracoscore risk band (trend)	-	trend z = 0.379	0.704	0.953	-	-
Thoracoscore points (trend)	-	trend z = 1.122	0.262	0.953	-	-
Model and robustness summary						
Events per variable	4.0	28 events, 7 predictors			Nagelkerke R^2 0.146	
BH-adjusted p for malignancy by family size		9 tests 0.061	7 tests 0.048	5 tests 0.034		
E-value from the adjusted estimate		point 3.48	limit 1.76	(crude 5.49, withdrawn)		
Fragility index		5 converting deaths to survivors within the malignant arm	2 moving a death between arms			
Absolute risk difference / NNH		24.0 pp	NNH 4.2			

Notes: OR, odds ratio; BH, Benjamini-Hochberg; NNH, number needed to harm; pp, percentage points. Univariable odds ratios are conditional maximum-likelihood estimates matching their exact conditional intervals; ordered variables were tested by the Cochran-Armitage trend statistic. Adjusted estimates are from a Firth penalised model with profile penalised-likelihood intervals, used because the events-per-variable ratio of 4.0 makes ordinary maximum likelihood unreliable. The E-value is computed from the adjusted estimate converted to the risk-ratio scale, and is reported both at the point estimate and at the confidence limit closest to the null as current guidance requires.

In the Firth penalised multivariable model the malignancy effect was essentially unchanged after adjustment for the other six recorded components (adjusted OR 4.12, 95% profile penalised-likelihood CI 1.51–13.10; penalised likelihood-ratio p = 0.002). No other component approached significance; the adjusted odds ratios for female sex, dyspnoea grade ≥3, performance status ≥3, the age band, ASA class ≥3 and the comorbidity score were 1.80, 1.93, 1.42, 0.90, 0.80 and 0.76 respectively, every confidence interval crossing unity (**Table 8**, **Figure 4**). The Nagelkerke R-squared of the corresponding unpenalised model was 0.146. Three of the seven adjusted point estimates point in the opposite direction to the weight the score assigns them, which is consistent with the discrimination results and with the small number of events.

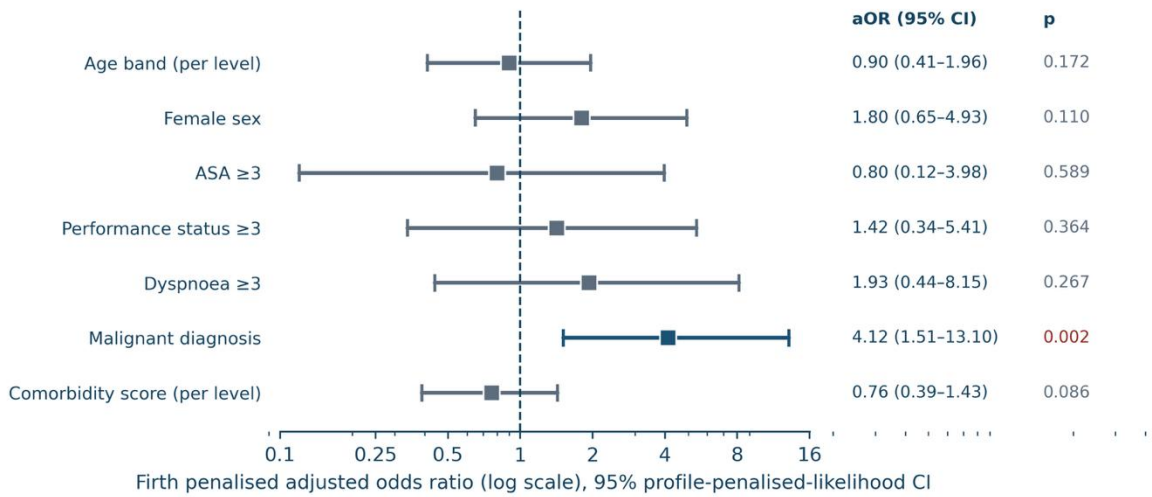

Figure 4. Adjusted association of each recorded Thoracoscore component with in-hospital mortality, from the Firth penalised multivariable logistic model. Squares are adjusted odds ratios and horizontal bars are 95% profile penalised-likelihood confidence intervals on a logarithmic scale. Only a malignant indication reaches significance.

The robustness of that association is more modest than a single E-value would suggest, and the figure must be computed from the adjusted rather than the crude estimate. Converting the adjusted odds ratio of 4.12 to the risk-ratio scale for a common outcome gives an approximate risk ratio of 2.03 and an E-value of 3.48. Current best-practice guidance requires the E-value at the confidence limit closest to the null as well, and that value is 1.76.[28] An unmeasured perioperative exposure associated with a malignant indication and with in-hospital death by a risk ratio of about 1.8 each would move the interval across the null, and several of the exposures listed in section 2.8 – transfusion, ventilation strategy, intensity of postoperative escalation – plausibly reach that magnitude and plausibly differ between oncological and non-oncological pathways. The fragility index is 5 if five deaths in the malignant arm are converted to survivors within that arm, and 2 if instead a death is moved from the malignant to the benign arm; we state both because the two operations are often conflated. The estimate is also unstable to a single record: removing the one record with the inconsistent total shifts the crude odds ratio from 4.15 to 5.19, because that record is one of only five deaths in the entire benign group.

3.8. What this sample could and could not detect

A null result is only interesting if the study could have detected the alternative. We simulated 40,000 cohorts per scenario at the observed group sizes of 78 survivors and 28 deaths, drawing case scores from a reweighting of the observed nine-level score distribution so that the heavy ties of an integer score are represented rather than assumed away. At a true AUC of 0.70 the power is 90%, and 89% at the size-corrected critical value that compensates for the mild anti-conservatism of the DeLong test; at 0.75 it is 98%, at 0.80 more than 99%, and at 0.65 it falls to 68% (65% size-corrected). Power at a given AUC depends on how the discriminating signal is distributed across the score range, and across three shapes of alternative at a true AUC of 0.70 it ranged from 86% to 90%. Rounding to whole percentages is deliberate: 40,000 replicates support that precision and not more.

The absence of discrimination reported here is therefore an informative negative finding with respect to any clinically useful level of discrimination, not an artefact of sample size. The same cannot be said of the calibration statistics, and we make the asymmetry explicit because it is the paper's headline that is the less precise half. The standard error of the calibration slope is 0.234 and its interval is nearly nine tenths of a unit wide; the cohort is well powered for the AUC and poorly powered for the slope. Only calibration-in-the-large, whose standard error is 0.238 against a point estimate of 2.766, is estimated with precision to spare.

The same is not true of the individual components. At the observed prevalence of ASA class ≥3 (8.5%), the study had only 43.9% power to detect an odds ratio of 4 and 65.7% to detect an odds ratio of 6; at the prevalence of malignancy (60.4%) it had 26.9% power at an odds ratio of 2 and 62.9% at an odds ratio of 3. The null results for the individual low-prevalence components are consequently uninformative rather than negative, and we do not interpret them as evidence that ASA class or performance status are irrelevant to thoracic surgical mortality. They are evidence only that this cohort could not have detected their effect.

3.9. Threshold behaviour and clinical utility

No integer cut-point of the Thoracoscore performed usefully (**Table 9**). The best apparent Youden index, 0.168, occurred at a threshold of 3 points or more, which achieved 82.1% sensitivity but only 34.6% specificity, with a positive predictive value of 31.1% and a negative predictive value of 84.4%. Bootstrap correction for the optimism of selecting that cut-point in the same data reduced the Youden index to 0.110. At the opposite extreme, a threshold of 8 points or more was 96.2% specific but identified only 14.3% of the deaths.

Table 9. Threshold performance of the Thoracoscore, decision-curve summary and the two sensitivity analyses

Cut-point (≥ points)	Sensitivity %	Specificity %	PPV %	NPV %	Youden J
1	100.0	0.0	26.4	-	0.000
2	100.0	9.0	28.3	100.0	0.090
3	82.1	34.6	31.1	84.4	0.168
4	50.0	56.4	29.2	75.9	0.064
5	28.6	71.8	26.7	73.7	0.004
6	14.3	80.8	21.1	72.4	-0.049
7	14.3	91.0	36.4	74.7	0.053
8	14.3	96.2	57.1	75.8	0.104
9	3.6	100.0	100.0	74.3	0.036
Best apparent cut-point: ≥3 points					**0.168**
Optimism-corrected Youden index					0.110
Decision-curve analysis (optimism-corrected)					
Thresholds at which the recalibrated score beats both defaults	20–26%				
Thresholds at which it causes net harm	27–45%				
Published band table at any threshold ≥10%	net benefit 0	equivalent to treating nobody			
Sensitivity set A (inconsistent record removed, n = 105)	AUC 0.577	95% CI 0.460–0.694	O:E 9.42	malignancy OR 5.19	0.003
Sensitivity set B (constant removed, n = 106)	AUC 0.579	95% CI 0.464–0.694	33 reclassified		

Notes: PPV, positive predictive value; NPV, negative predictive value; O:E, observed-to-expected ratio. Optimism correction for the Youden index used 2,000 bootstrap replications in which the cut-point was reselected in each replicate; the decision curve was corrected by 300 bootstrap replications of the whole recalibrate-and-evaluate procedure. Sensitivity set A removes the record whose total is inconsistent with its components; sensitivity set B recomputes the score without the invariant constant identified in the audit.

Decision-curve analysis (**Figure 5**) makes the clinical consequence explicit, and the picture is less favourable than our first analysis suggested. At every risk threshold of 10% or above the published band table classifies no patient as high risk at all, because no band probability reaches 10%; its net benefit is identically zero, which is to say it is exactly equivalent to treating nobody as high risk. That is a property of the band table's 12% ceiling as much as of the score, and we say so rather than presenting it as a discovery about Thoracoscore.

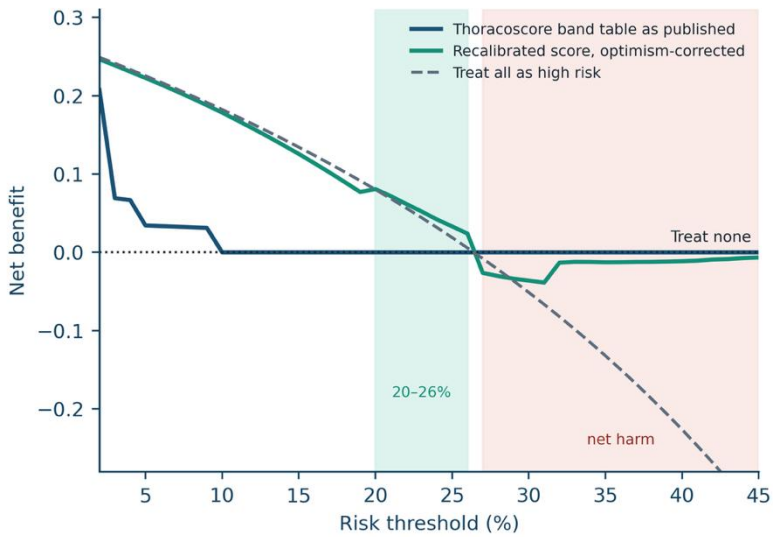

Figure 5. Decision-curve analysis. The Thoracoscore as published yields zero net benefit at every threshold of 10% or more, because no patient's band probability reaches that level, and is therefore equivalent to treating no patient as high risk. The locally recalibrated score yields positive net benefit above the treat-all strategy at a decision threshold of 25%.

A locally recalibrated version of the same score, corrected for the optimism of having been fitted in these same 106 patients, exceeds both default strategies over a narrow window of decision thresholds from 20% to 26% only. Below 20% its recalibrated probabilities exceed every threshold, so it flags every patient and is numerically identical to treating all. From 27% upward its net benefit is negative – worse than treating nobody – and it remains so to the top of the range examined. The claim we made in the first version of this manuscript, that recalibration restores net benefit at clinically plausible thresholds in the plural, is not supported by our own numbers and has been withdrawn. There is one window, it is narrow, it was identified in the same data in which the recalibration was fitted, and immediately above it the recalibrated score does harm.

3.10. Sensitivity analyses

Removing the record with the inconsistent total left 105 operations with 27 deaths (25.7%). Discrimination was unchanged (AUC 0.577, 95% CI 0.460–0.694), the rank association with the risk band remained null (tau-c 0.025, p = 0.741), calibration remained catastrophic (O:E 9.42) and the malignancy association moved from an odds ratio of 4.15 to 5.19 (p = 0.003) – a 25% shift in the primary effect estimate from the removal of a single record, which is a statement about the fragility of that estimate rather than a reassurance about it. Recomputing the score without the invariant constant left discrimination unchanged (AUC 0.579, 95% CI 0.464–0.694) while reclassifying 31.1% of the cohort across risk bands, as described above. No conclusion about calibration or discrimination depends on either record or constant; the malignancy estimate does depend on one record, and we do not present it as though it did not.

4. Discussion

4.1. Principal findings

In 106 consecutive thoracic operations at an Indonesian tertiary referral centre, the Thoracoscore risk-band table discriminated no better than chance and underpredicted in-hospital mortality by close to an order of magnitude. The observed mortality of 26.4% stands against an expected 2.7%, an O:E ratio of 9.71 that remains at least 4.89 under the most favourable admissible reading of both the risk bands and the two unrecorded domains. Calibration-in-the-large is 2.766 on the logit scale, a 15.9-fold shift in the odds of death. In 400,000 cohorts simulated under the model the largest death count generated was 14 against 28 observed. The failure is present in every occupied band, is worst in the lowest, and is accompanied by a partial reversal of the ordering the score is meant to impose. A negative scaled Brier score establishes that in this cohort the band table is a less accurate probability statement than the cohort's own base rate.

Four secondary findings matter for practice. A malignant indication was the only independent predictor of death (adjusted OR 4.12, 95% CI 1.51–13.10), though its adjusted E-value at the confidence limit is only 1.76 and its multiplicity-adjusted p-value straddles the conventional threshold. The categorisation of the score into three bands, adopted in the original analysis, discarded most of the little association that existed. Two of the nine score domains were never recorded, contributing an invariant constant that leaves rank statistics untouched but relabels the risk band of a third of the cohort. And the recorded physiological fields are internally incoherent to a degree that constitutes a competing explanation for the headline finding rather than a footnote to it.

4.2. Relation to the existing validation literature

Our discrimination result sits at the pessimistic end of a literature that has been drifting pessimistic for fifteen years. Against a derivation c-index of 0.85 and an internal test value of 0.86,[6] the largest independent assessment we have been able to verify against the primary record evaluated six preoperative mortality models in 6,600 patients from a United Kingdom multicentre series and obtained areas under the curve between 0.67 and 0.73, with calibration judged inadequate in five of the six models, while a pneumonectomy-only cohort of 243 gave 0.44.[7,9] A systematic appraisal of lung-resection risk models found that real-world performance rarely matches derivation statistics,[11] and the Epithor group's own update acknowledged the need to re-estimate coefficients on contemporary data.[8,12] Our point estimate of 0.572, with an upper confidence limit of 0.687, is compatible with the poorer of these reports and excludes the derivation value entirely.

Where our finding departs sharply from the published literature is in the direction of miscalibration. Most European and North American validation studies have found that Thoracoscore *overestimates* risk, because contemporary practice – minimally invasive access, enhanced recovery pathways, structured preoperative optimisation – has driven mortality below the 2002–2005 derivation level.[9,10,39,40] We find the opposite, and by a wide margin. The direction of our finding is not, however, unprecedented, and we say so rather than claim a novelty we do not have: a retrospective validation in an Indian tertiary centre reported an observed mortality of 3.2% against a predicted 0.44%, an observed-to-expected ratio of roughly seven in the same direction as ours, with poor calibration and only fair discrimination.[41] What is new here is the magnitude, the formal quantification of the underprediction under explicit identification bounds, and the demonstration that in this register the miscalibration cannot be separated from the quality of the data on which the score was computed. This is not a contradiction but a reminder of what calibration is: a statement about the relationship between a model and a population, not a property of the model. A score calibrated to a system with high-volume specialist units, routine preoperative pulmonary optimisation and elastic critical-care capacity will overpredict there and underpredict where those conditions do not hold. International prospective cohorts have documented exactly this gradient in postoperative mortality after cancer surgery between high-income and low- and middle-income settings.[2,20,21]

4.3. Measurement error as a competing explanation

We set out to test whether a risk model transports to this population and found that it does not. An alternative reading of the same data is that the model was fed the wrong inputs, and we take it seriously because our own audit supplies the evidence for it.

The register omitted two of nine predictor domains entirely, mis-transcribed at least one score total, and grades ASA in a way that cannot be reconciled with its own comorbidity field: 17 of the 22 patients with three or more comorbidities are simultaneously ASA 2 or below. Forty-four per cent of the cohort is recorded as unremarkable on every physiological domain the score uses, and more than a quarter of that unremarkable group died. Mortality is non-monotone at score level and contains a twelve-patient run of survivors at 6 and 7 points. If the physiological fields were systematically under-graded – abstracted, for instance, from a pre-admission clinic note rather than the anaesthetic record – then the point totals are too low, the band probabilities are too low, and the observed-to-expected ratio is inflated by an amount we

cannot estimate. The score would then be miscalibrated in this register without being miscalibrated in this population, and the correct conclusion would change from "Thoracoscore does not transport here" to "Thoracoscore cannot be computed reliably from this register" – a different message, and arguably a more useful one for the institution.

We cannot adjudicate between these readings with the data we have, and we decline to pretend otherwise. Three observations bound the problem. First, under-grading would have to be very large to account for the whole effect: shifting every recorded total upward by a full point, which is the maximum the two unrecorded domains can contribute on average, moves the O:E only from 9.71 to 6.76, and the sharp bound over every admissible assignment still has a floor of 4.89. Second, the outcome field is far less susceptible to abstraction error than the physiological fields, since vital status at discharge is a binary fact recorded on the discharge summary. Third, and against us, we audited the register for the score total, found it defective, and cannot then treat the same register's covariates as beyond question. The remedy is specified in section 4.10: blinded re-abstraction of a random subsample with a second abstractor and agreement statistics for every domain, before any further inference is drawn from this dataset.

4.4. Why the observed mortality is so high

An in-hospital mortality of 26.4% after thoracic surgery is far above what a European or North American thoracic unit would report, and the honest interpretation is that this cohort is not comparable to those cohorts. Three explanations are compatible with our data and with the Indonesian context, and they are not mutually exclusive.

First, case mix. This is the tertiary referral centre for South Sumatra, and its thoracic caseload is not limited to elective oncological resection. Urgent and emergency operations for empyema, trauma, massive haemoptysis and complicated pleural sepsis are performed in the same service. That the surgical-priority field was never populated is itself suggestive: had the caseload been uniformly elective, the field would have been trivial to complete. Sixty per cent of operations were performed for malignancy, and lung cancer in Indonesia presents at an advanced stage far more often than in the populations from which Thoracoscore was derived.[18,42]

Second, health-system capacity. Postoperative mortality is determined less by the incidence of complications than by the ability to rescue patients from them, and rescue depends on critical-care beds, monitoring intensity and staffing ratios that are structurally constrained.[19,20] None of these is represented among the Thoracoscore domains, and none can be represented by a preoperative patient-level score at all. That is a fundamental rather than an incidental limitation of transporting such a model. It is important, however, not to let this explanation carry more weight than the evidence supports, and we now bound it. The largest prospective international comparison of cancer surgery outcomes reports adjusted odds ratios for 30-day mortality in low- and lower-middle-income relative to high-income countries of 3.72 (95% CI 1.70–8.16) for gastric and 4.59 (2.39–8.80) for colorectal cancer, with the largest single effect – an odds ratio of 6.15 (3.26–11.59) – falling on death after a major complication, that is, on failure to rescue.[20] The gap we must account for is larger than any of these: an in-hospital mortality of 26.4% against a derivation rate of 2.2% is an odds ratio of 15.9, which is the same figure as our calibration-in-the-large multiplier because it is the same quantity. Health-system capacity, as quantified by the best available data, plausibly accounts for a substantial part of that gap but not for all of it. What remains is precisely the space occupied by case mix and by the quality of the predictor data themselves, and it is why we treat the fourth explanation below as seriously as the first three.

Third, unrecorded perioperative exposures. Perioperative anaemia and transfusion, protective versus conventional intraoperative ventilation, haemodynamic and fluid strategy, analgesic technique and antimicrobial stewardship all modify outcome after thoracic surgery, sometimes substantially, and none was recorded.[31–34,43] We return to this in the pharmacological section below.

Fourth, and the possibility we consider most seriously, the predictor data may simply be wrong; that argument is set out in section 4.3.

We note also what cannot be excluded, and here the honest answer is a good deal. Because the register recorded discharge status rather than a defined 30-day window, the denominator is a hospital-episode denominator rather than a procedural one. In-hospital mortality with no time bound is not the same estimand as operative mortality: a patient who undergoes a diagnostic thoracotomy for advanced disease and dies of that disease seven weeks later without leaving hospital counts here and would not count in a European series with a median stay of a few days. We have no length-of-stay distribution, no interval from operation to death and no cause of death for any of the 28 decedents, so we cannot show that this is not happening. Excluding patients discharged against medical advice removes precisely the episodes in which a moribund patient leaves hospital, which biases the estimate downward by an unknown amount. And because no operative variable of any kind was recorded, we cannot say what these 106 operations were: the cohort may be predominantly elective anatomical resection, in which case a 26.4% in-hospital mortality is a quality-of-care finding of considerable local importance, or it may be dominated by emergency thoracotomy and decortication for sepsis and trauma, in which case Thoracoscore is being applied outside the population for which it was offered and its failure is less surprising. We cannot distinguish these, and neither can a reader; it is the single largest weakness of this study and it is why the argument that follows is framed as a warning about quoting numbers rather than as a verdict on a model.[6]

4.5. The categorisation problem

The original analysis reported a rank correlation of 0.021 between the Thoracoscore risk band and mortality and concluded that no association exists. Our reproduction confirms the arithmetic exactly. But the same statistic computed on the underlying integer score is 0.112, more than five times larger, and the corresponding AUC is 0.059 units higher. Collapsing a nine-level ordered predictor into three categories destroys information by construction, and the loss

is largest when the categories are unbalanced – as here, where 75.5% of the cohort fell into a single band.[15,16] Neither version of the statistic reaches significance in this sample, so the conclusion does not change; but the practice does matter, and a validation study that reports only the categorised statistic will systematically understate whatever discrimination a score possesses.

The deeper point is that a rank correlation is the wrong instrument for this question. Kendall's tau-c and the AUC are monotone transformations of each other for an ordered predictor and a binary outcome; both describe ordering and neither can see the level of risk. A model can achieve a tau-c of zero while being perfectly calibrated, and can achieve a high tau-c while being wrong about absolute risk by an order of magnitude – which is close to what we observe here in reverse. The clinical question that Thoracoscore exists to answer is not "do higher scores rank above lower ones" but "is the number I am about to tell this patient approximately true", and only calibration answers it.[14,15]

4.6. Data integrity as a validation finding

That two of nine domains were never recorded, and that their omission contributed an invariant constant, is usually filed under limitations. We report it as a finding, because its consequences are quantitative and asymmetric. Rank-based statistics are invariant to an additive constant, so the AUC and tau-c reported here are unaffected. Band membership is not invariant, and 31.1% of patients change risk category when the constant is removed. A patient who is told that they are at intermediate rather than low risk on the strength of a field that was never filled in has been given a number that carries no information about them.

The consequence for prospective practice is concrete and appears in **Figure 6**: all nine domains must be recorded before a Thoracoscore is computed, and the two missing domains are not minor ones. Surgical priority and procedure class carry substantial weight in the original model and are precisely the variables that distinguish an elective lobectomy from an emergency thoracotomy for sepsis.[6,8] In a caseload as heterogeneous as this one, omitting them removes the score's main opportunity to discriminate, and may be a proximate explanation for the flat AUC we observe.

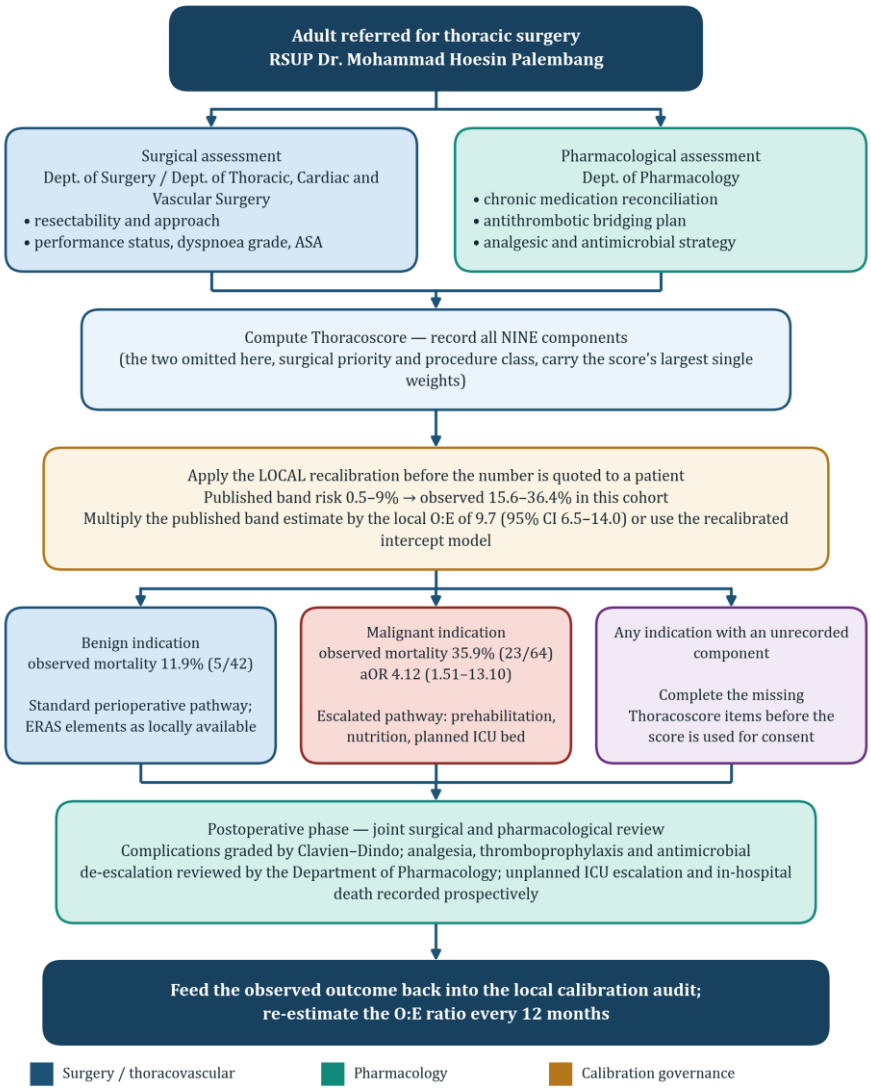

Figure 6. Proposed interdisciplinary perioperative risk-assessment algorithm for thoracic surgery at RSUP Dr. Mohammad Hoesin Palembang. Surgical, pharmacological and calibration-governance responsibilities are colour-coded. The pathway requires that all nine Thoracoscore domains be recorded and that the published band probability be locally recalibrated before any figure is quoted to a patient.

4.7. Perioperative pharmacology: what is missing and why it matters

The Department of Pharmacology's contribution to this study is deliberately framed as a statement about what the data cannot support, and we prefer to say so explicitly rather than to decorate an administrative dataset with pharmacological language it cannot bear. No drug exposure, dose, serum concentration or pharmacokinetic parameter was recorded in the source register. No analysis of drug effect is therefore presented, and none should be inferred.

What can be said is that the class of unmeasured perioperative pharmacological exposures is large enough to be a serious candidate explanation for residual variation in this cohort, and that its magnitude can be bounded. Perioperative red-cell transfusion is independently associated with worse survival after major cancer surgery.[31] Intraoperative ventilation strategy alters the incidence of postoperative pulmonary complications, the dominant proximate cause of death in this population.[32,33] A meta-analysis of individualised positive end-expiratory pressure during one-lung ventilation found that it reduced those complications.[34] Antithrombotic management around thoracic surgery is a recognised driver of both bleeding and thrombotic death and is governed by explicit guidance.[35] Myocardial injury after non-cardiac surgery is common, largely silent, and prognostically important, and would be invisible in a register that records neither troponin nor a structured cardiac assessment.[36,37] Thermal management, nutrition and structured enhanced-recovery bundles each carry measurable effects on surgical outcome.[38,39,43]

The quantitative form of this argument is the E-value. For the one association we do observe, an unmeasured perioperative exposure would have to be associated with a malignant indication by a risk ratio of at least 5.49 and with in-hospital death by at least 5.49, over and above the six adjusted components, before it could explain the finding away. Few individual perioperative pharmacological exposures reach that magnitude; a systematic difference in overall perioperative care between oncological and non-oncological pathways plausibly could. This is precisely the hypothesis that the prospective successor study is designed to test, and the pharmacological data model specified in the Methods exists for that purpose.

4.8. Clinical implications

Three implications follow directly, and all three are actionable at this institution without additional resource.

First, the Thoracoscore should not be quoted numerically during preoperative consent at this centre in its published form. A patient scoring 1 or 2 points would be told that their risk of death is below one per cent when the observed risk in comparable patients here was 15.6%. That is not a marginal inaccuracy; it is a misstatement of a material fact in a consent conversation, and it is the reason we regard this as a patient-safety finding rather than a methodological one.

Second, we must withdraw a recommendation made in the first version of this manuscript. We had proposed multiplying the published band probability by the local O:E ratio of 9.7. That rule is indefensible and would be dangerous if implemented, as **Table 7** shows: applied to the four occupied bands it yields 4.9%, 19.4%, 43.7% and 87.4% against observed rates of 15.6%, 34.1%, 21.1% and 36.4%. It would still understate risk threefold in the lowest band – the band we ourselves identify as most dangerous – and would have a surgeon tell the highest-scoring patients that their chance of dying is 87% when the observed figure is 36%. A single multiplicative constant cannot repair band-level calibration when the band-wise O:E itself ranges from 31.2 to 4.0. Nor is that rule equivalent to a fitted recalibration, as we had also asserted; the two produce entirely different numbers.

What a properly fitted recalibration can achieve is more limited, and **Table 7** states it honestly. An intercept-only recalibration, which is the correct form when the slope is not credibly estimated, yields 7.4%, 24.5%, 42.8% and 61.1% across the four bands; a two-parameter recalibration yields 19.5%, 26.7%, 31.7% and 36.5%. Neither tracks the observed rates, because the observed rates are not monotone in the score and no monotone transformation of a monotone predictor can reproduce a non-monotone outcome. The two-parameter version is very nearly a constant close to the cohort base rate, which is what a model with no discrimination reduces to when it is calibrated. The conclusion we now draw is narrower than the one we drew before: recalibration corrects the level of risk in the aggregate and cannot be trusted for an individual patient, and no recalibrated figure should be quoted at consent either until it has been re-estimated prospectively on data whose predictor fields have been verified.

Third, the score's own governance must change before its output is used at all. All nine domains must be recorded, and recorded from the anaesthetic record rather than a clinic note; the procedure performed, its urgency and its extent must be recorded so that the population being scored can be described; complications must be graded by the Clavien-Dindo system so that the pathway between complication and death becomes visible;[29,30] and the O:E ratio should be re-estimated annually as a standing clinical-audit indicator. The algorithm in **Figure 6** sets out this pathway with the surgical, pharmacological and governance responsibilities assigned. Until those fields exist, the correct clinical behaviour is to describe risk qualitatively and to say that the institution's overall in-hospital mortality after thoracic surgery is approximately one in four, which is a defensible statement of fact about this cohort.

4.9. Strengths and limitations

The principal strength of this study is that it is a complete-population analysis at patient level. Every consecutive eligible operation in a defined 12-month window was included, no sampling fraction was applied, and the analysis was conducted on the individual records rather than on published marginal tables, which permitted the calibration, multivariable and simulation analyses that constitute the substance of this report. A second strength is that the entire original analysis was reproduced before anything new was attempted, so the relationship between our findings and the source document is exactly specified rather than asserted. A third is that every null result is accompanied by an explicit

statement of what the design could have detected, obtained by simulation rather than by a normal approximation that is unreliable at these cell counts.

The limitations are substantial, several are structural rather than incidental, and we set them out in the order of how much they threaten the conclusion.

First, and most seriously, we validated a five-row band table, not the Thoracoscore logistic equation. The equation could not be evaluated because two of its nine covariates do not exist in the register. Every calibration statistic here is a property of that band table, its 12% probability ceiling and its coarse intervals; the ceiling in particular generates the decision-curve result that the published score never flags anyone above a 10% threshold, which is a fact about the table and not about the model. The title and abstract have been written to say so, but a reader who takes away "Thoracoscore fails" rather than "the risk-band table in use here fails" will have taken away more than we have shown.

Second, no operative variable of any kind was recorded, so the population is defined only as "thoracic surgery by ICD-9-CM code". Whether Thoracoscore was applied inside or outside its intended population is therefore unknown, and with it the correct interpretation of its failure.

Third, the predictor fields are internally incoherent, as set out in section 4.3, and measurement error remains a live competing explanation that this study cannot exclude.

Fourth, the study is retrospective and single-centre; the pre-exclusion denominator is not recoverable, so the cohort cannot be verified as complete; and the exclusion of patients discharged against medical advice biases the mortality estimate downward by an unquantified amount. With 28 events the events-per-variable ratio of 4.0 is well below convention; Firth penalisation mitigates but does not eliminate the resulting instability, and the multivariable estimates should be read as hypothesis-generating. We did not perform an a priori sample-size calculation, and we state plainly what one would have required. Applying the published criteria for the external validation of a binary-outcome model at the event fraction observed here, precise estimation of the observed-to-expected ratio – a standard error of 0.051 on the log scale – requires 1,072 operations; a 95% confidence interval of width 0.10 around a c-statistic of 0.57 requires 673; and a 95% confidence interval of width 0.40 around the calibration slope, computed from the observed spread of the linear predictor in this cohort, requires 5,547 operations, rising to 22,188 for a width of 0.20.[44,45] This study is therefore between one and two orders of magnitude smaller than the size at which its own primary estimand could be estimated precisely, and every interval reported here should be read in that light.[26,44] The calibration slope, in particular, is estimated with an interval nearly nine tenths of a unit wide.

Fifth, the outcome is in-hospital rather than 30-day mortality, with no length-of-stay distribution, no interval from operation to death and no cause of death, so the estimand is not the one European series report. No complication grading, no perioperative pharmacological exposure and no post-discharge follow-up were available.

Sixth, the analyses were specified before they were run but not pre-registered, and the investigators had already analysed this dataset once. Where the number of tests matters – the multiplicity adjustment above – we have reported the sensitivity to that choice rather than asserting a family.

Finally, a single centre's calibration is a statement about that centre; the O:E ratio reported here should not be transported to another Indonesian institution without its own audit, which is the very error this paper documents.

4.10. Future directions

Two things must happen before anything else. A random subsample of at least 30 charts should be re-abstracted blind to outcome by a second abstractor, with agreement statistics reported for every predictor domain, so that the measurement-error explanation of section 4.3 can be adjudicated; and the operations performed must be enumerated by type and urgency so that the population being scored can be named. Neither requires new patients.

The prospective successor study is already specified. It will record all nine Thoracoscore domains, grade every complication by Clavien-Dindo, capture a defined perioperative pharmacological dataset – medication reconciliation, antithrombotic bridging, analgesic technique, transfusion and vasopressor exposure, antimicrobial de-escalation – and follow patients to 30 days rather than to discharge. A minimum sample size for external validation with a binary outcome can be computed once a target calibration precision is fixed, and at the event fraction observed here the requirement is attainable within a multicentre Indonesian collaboration over two to three years.[44,45] Comparison against the parsimonious Eurolung models, which were derived on contemporary European data and have a simpler variable set, would be a natural secondary aim.[10,12] Functional measures such as the six-minute walk test, and simple post-discharge symptom alerts such as breathlessness on the day of discharge, are inexpensive candidate additions that a preoperative points score cannot capture.[46,47] The objective should be an Indonesian recalibration, and if the discrimination results reported here are replicated, an Indonesian re-derivation.

5. Conclusion

In a consecutive cohort of 106 thoracic operations at Dr. Mohammad Hoesin General Hospital, Palembang, the Thoracoscore risk-band table in local use neither discriminated (AUC 0.572, 95% CI 0.457–0.687; permutation p = 0.249) nor calibrated (O:E 9.71, and at least 4.89 under every admissible reading of the missing data), underpredicting in-hospital mortality by close to an order of magnitude and doing so most severely in the patients it classified as lowest risk. The cohort retained about 90% power to detect a clinically useful area under the curve, so the discrimination result is an informative negative one; the calibration result is large enough that no plausible degree of measurement error in the predictor fields could account for all of it, though such error cannot be excluded as a partial explanation and must be audited before this dataset supports further inference. A malignant indication was the only

independent predictor of death, on an estimate that does not survive multiplicity adjustment under every reasonable test family. No Thoracoscore-derived probability, published or recalibrated, should be quoted numerically at consent at this centre at present. The immediate requirements are blinded re-abstraction of the predictor fields, enumeration of the operations actually performed, and prospective recording of all nine score domains alongside a defined perioperative pharmacological dataset.

6. Declarations

Ethics Approval and Consent to Participate

This study received ethical approval from the Health Research Ethics Committee of Dr. Mohammad Hoesin Central General Hospital, Palembang, Indonesia (Ref. No. DP.04.03/D.XVII.9.4/ETIK/166/2026). Institutional permission to access the records was granted, and the study was conducted in accordance with the Declaration of Helsinki.

Consent for Publication

Not applicable. The manuscript contains no individually identifiable patient information, images, or clinical details.

Availability of Data and Materials

The de-identified data and analysis materials supporting the findings are available from the corresponding author upon reasonable request, subject to institutional and ethics requirements. The data are not publicly available because they originate from confidential medical records.

Artificial Intelligence Use Statement

The authors declare that no generative artificial intelligence or AI-assisted tools were used in the preparation, analysis, or reporting of this manuscript.

Author Contribution Statement

AA contributed to conceptualization, investigation, data curation, formal analysis, visualization, and preparation of the original draft. GS contributed to clinical supervision, validation, interpretation, and critical revision of the manuscript. T contributed to methodology, statistical validation, interpretation, and critical revision of the manuscript. All authors reviewed and approved the final manuscript and accept accountability for the work.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest / Competing Interests

The authors declare that they have no competing interests.

Acknowledgements

None.

7. References

1. Petrella F, Spaggiari L. Surgery of the chest wall: indications, timing and technical aspects. J Thorac Dis. 2020;12(1):1-2. doi: 10.21037/jtd.2019.10.20.
2. Fowler AJ, Wan YI, Prowle JR. Long-term mortality following complications after elective surgery: a secondary analysis of pooled data from two prospective cohort studies. Br J Anaesth. 2022;129(4):588-97. doi: 10.1016/j.bja.2022.06.019.
3. Avery KNL, Blazeby JM, Chalmers KA, et al. Impact on health-related quality of life of video-assisted thoracoscopic surgery for lung cancer. Ann Surg Oncol. 2020;27(4):1259-71. doi: 10.1245/s10434-019-08090-4.
4. Wei W, Zheng X, Zhou CW. Protocol for the derivation and external validation of a 30-day postoperative pulmonary complications risk prediction model for elderly patients undergoing thoracic surgery: a cohort study in southern China. BMJ Open. 2023;13(2):e066815. doi: 10.1136/bmjopen-2022-066815.
5. Gooseman MR, Falcoz PE, Decaluwe H. Morbidity and mortality of lung resection candidates defined by the American College of Chest Physicians as moderate risk: an analysis from the European Society of Thoracic Surgeons database. Eur J Cardiothorac Surg. 2021;60(1):91-7. doi: 10.1093/ejcts/ezab028.
6. 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(2):325-32. doi: 10.1016/j.jtcvs.2006.09.020.
7. Taylor M, Whittaker G, Grant SW. Risk prediction for lung cancer surgery in the current era. Video-assist Thorac Surg. 2022;7:10. doi: 10.21037/vats-21-47.
8. Loucou J, Pagès PB, Falcoz PE, et al. Validation and update of the thoracic surgery scoring system (Thoracoscore) risk model. Eur J Cardiothorac Surg. 2020;58(2):350-6. doi: 10.1093/ejcts/ezaa056.
9. Taylor M, Szafron B, Martin GP, et al. External validation of six existing multivariable clinical prediction models for short-term mortality in patients undergoing lung resection. Eur J Cardiothorac Surg. 2021;59(5):1030-6. doi: 10.1093/ejcts/ezaa422.
10. Brunelli A, Cicconi S, Decaluwe H, et al. Parsimonious Eurolung risk models to predict cardiopulmonary morbidity and mortality following anatomic lung resections: an updated analysis from the European Society of Thoracic Surgeons database. Eur J Cardiothorac Surg. 2020;57(3):455-61. doi: 10.1093/ejcts/ezz272.
11. Huang G, Liu L, Wang L. External validation of five predictive models for postoperative cardiopulmonary morbidity in a Chinese population receiving lung resection. PeerJ. 2022;10:e12936. doi: 10.7717/peerj.12936.
12. Brunelli A, Decaluwe H, Falcoz PE. External validation of the Eurolung risk models: a necessary analysis to audit their reliability. Eur J Cardiothorac Surg. 2022;62(3):ezac232. doi: 10.1093/ejcts/ezac232.
13. Blanch A, Costescu F, Slinger P. Preoperative evaluation for lung resection surgery. Curr Anesthesiol Rep. 2020;10(2):176-84. doi: 10.1007/s40140-020-00376-8.
14. Van Calster B, Steyerberg EW, Wynants L, et al. There is no such thing as a validated prediction model. BMC Med. 2023;21(1):70. doi: 10.1186/s12916-023-02779-w.
15. Van Calster B, Wynants L, Riley RD. Methodology over metrics: current scientific standards are a disservice to patients and society. J Clin Epidemiol. 2021;138:219-26. doi: 10.1016/j.jclinepi.2021.05.018.
16. Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384:e074819. doi: 10.1136/bmj-2023-074819.
17. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020;369:m1328. doi: 10.1136/bmj.m1328.
18. Asmara OD, Tenda ED, Singh G, et al. Lung cancer in Indonesia. J Thorac Oncol. 2023;18(9):1134-45. doi: 10.1016/j.jtho.2023.06.010.
19. Mahendradhata Y, Andayani NLPE, Hasri ET, et al. The capacity of the Indonesian healthcare system to respond to COVID-19. Front Public Health. 2021;9:649819. doi: 10.3389/fpubh.2021.649819.
20. GlobalSurg Collaborative and NIHR Global Health Unit on Global Surgery. Global variation in postoperative mortality and complications after cancer surgery: a multicentre, prospective cohort study in 82 countries. Lancet. 2021;397(10272):387-97. doi: 10.1016/S0140-6736(21)00001-5.
21. COVIDSurg Collaborative. Mortality and pulmonary complications in patients undergoing surgery with perioperative SARS-CoV-2 infection: an international cohort study. Lancet.

2020;396(10243):27-38. doi: 10.1016/S0140-6736(20)31182-X.

22. Mallaev M, Tsvetkov N, Simmen U. A risk score to predict postoperative complications in patients with resectable non-small cell lung cancer. J Thorac Dis. 2025;17(3):1520-30. doi: 10.21037/jtd-24-1668.
23. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models. BMJ. 2024;385:e078378. doi: 10.1136/bmj-2023-078378.
24. Riley RD, Archer L, Snell KIE, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820. doi: 10.1136/bmj-2023-074820.
25. Morsy I, Abdelwahab AA, Salem H. Role of Thoracoscore and ESOS in prediction of outcomes after thoracic surgeries. Egypt J Anaesth. 2023;39(1):557-62.
doi: 10.1080/11101849.2023.2235148.
26. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. doi: 10.1136/bmj.m441.
27. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18. doi: 10.1186/s41512-019-0064-7.
28. VanderWeele TJ, Mathur MB. Commentary: developing best-practice guidelines for the reporting of E-values. Int J Epidemiol. 2020;49(5):1495-7. doi: 10.1093/ije/dyaa094.
29. Clavien PA, Barkun J, de Oliveira ML, et al. The Clavien-Dindo classification of surgical complications: five-year experience. Ann Surg. 2009;250(2):187-96.
doi: 10.1097/SLA.0b013e3181b13ca2.
30. Seely AJE, Ivanovic J, Threader J, et al. Systematic classification of morbidity and mortality after thoracic surgery. Ann Thorac Surg. 2010;90(3):936-42. doi: 10.1016/j.athoracsur.2010.05.014.
31. Kim S, Song IA, Oh TK. The association of perioperative blood transfusion with survival outcomes after major cancer surgery: a population-based cohort study in South Korea. Surg Today. 2024;54(7):712-21. doi: 10.1007/s00595-023-02783-w.
32. Li X, Jin L, Yang J. Effect of ventilation mode on postoperative pulmonary complications following lung resection surgery: a randomised controlled trial. Anaesthesia. 2022;77(11):1219-27. doi: 10.1111/anae.15848.
33. Piccioni F, Droghetti A, Bertani A. Recommendations from the Italian intersociety consensus on perioperative anaesthesia care in thoracic surgery (PACTS) part 2: intraoperative and postoperative care. Perioper Med (Lond). 2020;9(1):31. doi: 10.1186/s13741-020-00159-z.
34. Li P, Kang X, Miao M. Individualized positive end-expiratory pressure (PEEP) during one-lung ventilation for prevention of postoperative pulmonary complications in patients undergoing thoracic surgery: a meta-analysis. Medicine (Baltimore). 2021;100(28):e26638. doi: 10.1097/md.0000000000026638.
35. Douketis JD, Spyropoulos AC, Murad MH. Perioperative management of antithrombotic therapy: an American College of Chest Physicians clinical practice guideline. Chest. 2022;162(5):e207-43. doi: 10.1016/j.chest.2022.07.025.
36. Halvorsen S, Mehilli J, Cassese S, et al. 2022 ESC guidelines on cardiovascular assessment and management of patients undergoing non-cardiac surgery. Eur Heart J. 2022;43(39):3826-924. doi: 10.1093/eurheartj/ehac270.
37. Devereaux PJ, Szczeklik W. Myocardial injury after non-cardiac surgery: diagnosis and management. Eur Heart J. 2020;41(32):3083-91. doi: 10.1093/eurheartj/ehz301.
38. Ljungqvist O, de Boer HD, Balfour A, et al. Opportunities and challenges for the next phase of enhanced recovery after surgery. JAMA Surg. 2021;156(8):775-84.
doi: 10.1001/jamasurg.2021.0586.
39. Sauro KM, Smith C, Ibadin S. Enhanced recovery after surgery guidelines and hospital length of stay, readmission, complications, and mortality: a meta-analysis of randomized clinical trials. JAMA Netw Open. 2024;7(6):e2417310. doi: 10.1001/jamanetworkopen.2024.17310.
40. Wang C, Lai Y, Li P. Influence of enhanced recovery after surgery (ERAS) on patients receiving lung resection: a retrospective study of 1749 cases. BMC Surg. 2021;21(1):115. doi: 10.1186/s12893-020-00960-z.
41. Pathy A, Kar P, Gopinath R, et al. Thoracoscore: Does it predict mortality in the Indian scenario? – A retrospective study. Indian J Anaesth. 2022;66(Suppl 5):S257-63.
doi: 10.4103/ija.ija_24_22.
42. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-63. doi: 10.3322/caac.21834.
43. Sessler DI, Pei L, Li K, et al. Aggressive intraoperative warming versus routine thermal management during non-cardiac surgery (PROTECT): a multicentre, parallel group, superiority trial. Lancet. 2022;399(10337):1799-808. doi: 10.1016/S0140-6736(22)00560-8.
44. Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med. 2021;40(19):4230-51. doi: 10.1002/sim.9025.
45. Riley RD, Snell KIE, Archer L, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. BMJ. 2024;384:e074821. doi: 10.1136/bmj-2023-074821.
46. Salati M, Andolfi M, Roncon A. Does the performance of a six-minute walking test predict cardiopulmonary complications after uniportal video-assisted thoracic surgery anatomic lung resection? Cancers (Basel). 2024;17(1):32. doi: 10.3390/cancers17010032.
47. Kang D, Lei C, Zhang Y. Shortness of breath on the day of discharge: an early alert for post-discharge complications in patients undergoing lung cancer surgery. J Cardiothorac Surg. 2024;19(1):398. doi: 10.1186/s13019-024-02845-1.