Radiology

ORIGINAL RESEARCH • THORACIC IMAGING

Clinical Utility of a CT-based AI Prognostic Model for Segmentectomy in Non–Small Cell Lung Cancer

Kwon Joong Na, MD • Young Tae Kim, MD, PhD • Jin Mo Goo, MD, PhD • Hyungjin Kim, MD, PhD

From the Department of Thoracic and Cardiovascular Surgery (K.J.N., Y.T.K.) and Department of Radiology (J.M.G., H.K.), Seoul National University Hospital and College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea; Seoul National University Cancer Research Institute, Seoul National University College of Medicine, Seoul, Korea (K.J.N., Y.T.K., J.M.G.); and Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul, Korea (J.M.G.). Received July 12, 2023; revision requested August 11; revision received February 27, 2024; accepted March 1. **Address correspondence to** H.K. (email: *khj.snuh@gmail.com*).

Supported by the Seoul National University Hospital Research Fund (03-2022-2170) and a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00207978). However, the funders had no role in the study design; in the collection, analysis, and interpretation of the data; in the writing of the report; and in the decision to submit the article for publication.

Conflicts of interest are listed at the end of this article.

Radiology 2024; 311(1):e231793 • https://doi.org/10.1148/radiol.231793 • Content codes: CH AI CT

Background: Currently, no tool exists for risk stratification in patients undergoing segmentectomy for non–small cell lung cancer (NSCLC).

Purpose: To develop and validate a deep learning (DL) prognostic model using preoperative CT scans and clinical and radiologic information for risk stratification in patients with clinical stage IA NSCLC undergoing segmentectomy.

Materials and Methods: In this single-center retrospective study, transfer learning of a pretrained model was performed for survival prediction in patients with clinical stage IA NSCLC who underwent lobectomy from January 2008 to March 2017. The internal set was divided into training, validation, and testing sets based on the assignments from the pretraining set. The model was tested on an independent test set of patients with clinical stage IA NSCLC who underwent segmentectomy from January 2010 to December 2017. Its prognostic performance was analyzed using the time-dependent area under the receiver operating characteristic curve (AUC), sensitivity, and specificity for freedom from recurrence (FFR) at 2 and 4 years and lung cancer–specific survival and overall survival at 4 and 6 years. The model sensitivity and specificity were compared with those of the Japan Clinical Oncology Group (JCOG) eligibility criteria for sublobar resection.

Results: The pretraining set included 1756 patients. Transfer learning was performed in an internal set of 730 patients (median age, 63 years [IQR, 56–70 years]; 366 male), and the segmentectomy test set included 222 patients (median age, 65 years [IQR, 58–71 years]; 114 male). The model performance for 2-year FFR was as follows: AUC, 0.86 (95% CI: 0.76, 0.96); sensitivity, 87.4% (7.17 of 8.21 patients; 95% CI: 59.4, 100); and specificity, 66.7% (136 of 204 patients; 95% CI: 60.2, 72.8). The model showed higher sensitivity for FFR than the JCOG criteria (87.4% vs 37.6% [3.08 of 8.21 patients], $P = .02$), with similar specificity.

Conclusion: The CT-based DL model identified patients at high risk among those with clinical stage IA NSCLC who underwent segmentectomy, outperforming the JCOG criteria.

Supplemental material is available for this article.

Lobectomy has been the standard surgical approach for early-stage lung cancer since a landmark clinical trial in 1995 (1). However, two recent phase 3 randomized clinical trials demonstrated promising results, suggesting that sublobar resection might be accepted as the new standard of treatment in a select subcategory of small peripheral non–small cell lung cancer (NSCLC) tumors (2,3). Briefly, the Cancer and Leukemia Group B (CALGB) 140503 trial (ClinicalTrials.gov registration no. NCT00499330) reported that sublobar resection (wedge resection or segmentectomy) was noninferior to lobectomy for disease-free survival, with similar overall survival (OS) (2). In the Japan Clinical Oncology Group (JCOG) 0802 trial (University Hospital Medical Information Network Clinical Trials Registry no. UMIN000002317), both superiority and noninferiority in OS were reported in the segmentectomy group compared with the lobectomy group (3).

In the JCOG0802 trial (3), 67 of 552 patients (12.1%) who underwent segmentectomy experienced disease recurrence, while in the CALGB140503 trial (2), 102 of 336 patients (30.4%) who underwent sublobar resection had recurrence. The JCOG0802 trial also noted more locoregional relapses in patients who had undergone segmentectomy (11%) compared with those who had undergone lobectomy (5%, $P = .002$) (3). However, there is currently no tool for risk stratification of patients undergoing sublobar resection.

Dimensional measurements of lung nodules on CT scans alone may not be sufficient for prognostication. In fact, risk model–based clinical decision-making is a better approach than simply using one or two clinical variables (4). For example, guidelines recommend model-based risk calculation to determine the management strategy for incidentally detected lung nodules (5,6). In addition, a prediction model may aid further fine-tuning or optimization of clinical trial eligibility criteria and help exclude individuals at high risk for segmentectomy even if they meet the criteria (7).

Abbreviations

AUC = area under the receiver operating characteristic curve, CALGB = Cancer and Leukemia Group B, DL = deep learning, FFR = freedom from recurrence, JCOG = Japan Clinical Oncology Group, LCSS = lung cancer–specific survival, NSCLC = non–small cell lung cancer, OS = overall survival

Summary

A CT-based deep learning model identified individuals at high risk among patients with clinical stage IA non–small cell lung cancer who underwent segmentectomy.

Key Results

- In this retrospective study of 222 patients who underwent segmentectomy for clinical stage IA non–small cell lung cancer, a CT-based deep learning (DL) model showed a time-dependent area under the receiver operating characteristic curve of 0.86, sensitivity of 87.4%, and specificity of 66.7% for recurrence within 2 years.
- The model showed higher sensitivity for 2-year recurrence than the Japan Clinical Oncology Group eligibility criteria for sublobar resection (87.4% vs 37.6%, $P = .02$).

Thus, the aim of this study was to develop and validate a deep learning (DL) model using both preoperative CT scans and clinical and radiologic information for prognostication and risk stratification in a retrospective cohort of patients who underwent segmentectomy for clinical stage IA NSCLC.

Materials and Methods

This retrospective study was approved by the institutional review board of Seoul National University Hospital. The requirement for written informed consent was waived because data were analyzed retrospectively and anonymously. Figure 1 presents a schematic illustration of the overall study design.

Study Patients

The DL model was trained in two stages, pretraining and transfer learning. The pretraining set included patients with any pathologic stage (ie, any tumor size and lymph node involvement, but without metastasis, confirmed at pathologic examination [pTanyNanyM0]) primary NSCLC (adenocarcinoma, squamous cell carcinoma, or large cell carcinoma) who underwent lobectomy without neoadjuvant therapy from January 2008 to March 2017 at Seoul National University Hospital, a tertiary referral center. Exclusion criteria are detailed in Appendix S1. Transfer learning was applied to a subset of the pretraining set, specifically in patients with clinical stage IA NSCLC (hereafter referred to as the internal set).

The independent segmentectomy test set included patients who underwent segmentectomy for clinical stage IA (according to the eighth edition staging system) NSCLC of any histologic type from January 2010 to December 2017 at the same hospital (8). Patients who underwent surgery for lung cancer within the past 5 years or who had recurrent lung cancer, a positive clinical N category, a forced expiratory volume in the first second of expiration less than 60%, or a nonmeasurable primary tumor were excluded. Patients who overlapped with the internal set due to metachronous lung cancer and those with clinical stage 0 NSCLC were also excluded.

Approximately 60% of patients (approximately 1200 of 1978 patients) in these data sets have been reported previously (9–16). However, none of the prior studies dealt with risk stratification in patients undergoing sublobar resection. The DL model in this study has not been reported or validated previously.

Model Pretraining and Transfer Learning

The pretraining set was split randomly for training, validation, and testing at a ratio of 6:2:2. The DL model was pretrained to predict visceral pleural invasion, lymphovascular invasion, lymph node metastasis, and OS through multitask learning using three-dimensional CT tumor patches. This approach was used to promote survival prediction by instilling the domain knowledge, including the results of the JCOG0201 studies, which were the basis of the eligibility criteria in the JCOG0802 and West Japan Oncology Group 4607L trial (3,17–22). The internal set was divided based on the assignments from the pretraining set, with the ratio of the internal training set (for transfer learning), internal validation set, and internal testing set being 6:1.8:2.2. Transfer learning of the cumulative OS probability in the internal set was completed using additional information that included age, sex, solid portion size on CT scans, and the consolidation to tumor ratio (3,23).

The pretraining and transfer learning of cumulative OS were performed using a discrete-time survival model for four time intervals (0–24, 24–48, 48–72, and 72–165 months) (24). The cumulative risk of death (hereafter referred to as the DL-driven risk score) as a percentage was defined as (1 – cumulative survival probability to each time point) × 100. Thus, DL-driven 2-year, 4-year, and 6-year risk scores were obtained. A higher score indicates greater risk. The code is available at *https://github.com/chestrad/segmentectomy*. Tumor annotation, data preprocessing, model development, visualization, and CT acquisition are detailed in Appendix S1.

Study Outcomes

The study outcomes were freedom from recurrence (FFR), lung cancer–specific survival (LCSS), and OS (25,26). FFR was measured from surgery to the first recurrence and/or distant metastasis, confirmed at imaging (CT or PET) or histologic examination, with censoring at the last follow-up date. LCSS was the interval from the date of surgery to the date of death caused by the same lung cancer. If patients died of an indeterminate cause, the censoring time was set at the last follow-up date. OS was calculated from the date of surgery to the date of death by any cause. Survival time was censored at June 8, 2022, for the pretraining set (including the internal set) and June 20, 2023, for the segmentectomy test set. Survival status and the date of death were acquired from the database of the Ministry of Interior and Safety, Republic of Korea.

OS was obtained for both the internal and segmentectomy test sets, but FFR and LCSS were acquired only for the segmentectomy test set. For the tertiary center, the standard follow-up protocol involved acquisition of CT scans at intervals of 6–12

months during the initial 2 years after surgery, followed by annual observations.

Statistical Analysis

The prognostic performance of the DL-driven risk scores was analyzed in the internal test set and the independent segmentectomy test set. The prognostic discrimination was evaluated using the time-dependent area under the receiver operating characteristic curve (AUC) at three time points (ie, 2, 4, and 6 years) (27). Specifically, FFR at 2 and 4 years, LCSS at 4 and 6 years, and OS at 4 and 6 years were analyzed. To calculate the time-dependent sensitivity and specificity with 95% bootstrap CIs (5000 replications) for each outcome (27), cutoffs for the DL-driven 2-year, 4-year, and 6-year risk scores, respectively, were determined empirically as the median values in the internal validation set. The cutoff for each DL risk score remained consistent regardless of the study outcome. Kaplan-Meier survival curves stratified according to the dichotomized DL risk scores (ie, high vs low risk) were plotted.

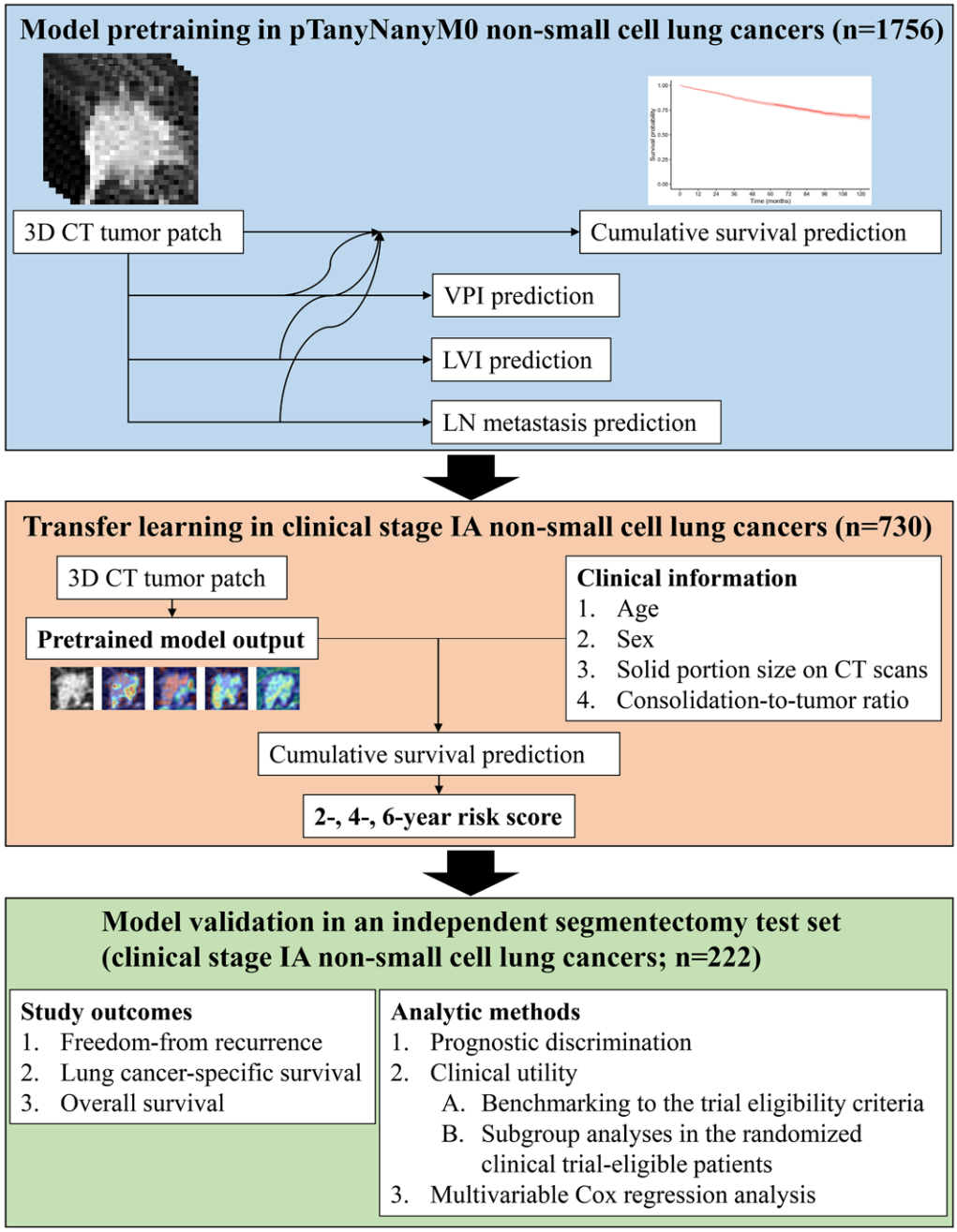

Figure 1: Schematic shows the overall study design. The pretraining set included patients with non–small cell lung cancer with any tumor size and lymph node involvement but without metastasis, confirmed at pathologic examination (pTanyNanyM0). LN = lymph node, LVI = lymphovascular invasion, 3D = three-dimensional, VPI = visceral pleural invasion.

The clinical utility of the DL risk score in the segmentectomy test set was evaluated in two ways as follows: *(a)* benchmarking to the clinical trial eligibility criteria and *(b)* conducting subgroup analyses in the trial-eligible patients. First, the sensitivity and specificity of DL-based risk grouping were compared with those of the clinical trial eligibility criteria. The JCOG eligibility criteria were defined as the union of JCOG0802, JCOG1211 (University Hospital Medical Information Network Clinical Trials Registry no. UMIN000011819), and JCOG0804 (no. UMIN000002008) criteria as follows (3,23). JCOG0802 was a randomized controlled noninferiority trial of segmentectomy versus lobectomy in patients with clinical stage IA NSCLC (total tumor size ≤2 cm and consolidation to tumor ratio >0.5) (3), and JCOG1211 was a single-arm trial for segmentectomy (total tumor size ≤3 cm and consolidation to tumor ratio ≤0.5, except tumors ≤2 cm with a consolidation to tumor ratio ≤0.25) (23). JCOG0804 was a single-arm trial for wedge resection or segmentectomy (tumors ≤2 cm and consolidation to tumor ratio ≤0.25) (28). The eligibility criterion of the CALGB140503 trial was also evaluated (tumors with a solid component ≤2 cm) (2). Patients with pure ground-glass nodules or pathologically confirmed N1 or N2 disease were

Table 1: Patient and Tumor Characteristics

Characteristic	Internal Set (*n* = 730)*	Segmentectomy Test Set (*n* = 222)	*P* Value
Age (y)†	63 (56–70) [29–86]	65 (58–71) [31–85]	.008
Sex			.81
M	366 (50.1)	114 (51.4)	
F	364 (49.9)	108 (48.6)	
History of malignancy other than lung cancer			
Yes	155 (21.3)	NA	
No	574 (78.7)	NA	
Family history of lung cancer			
Yes	42 (5.8)	NA	
No	686 (94.2)	NA	
Smoking history			.54
Never	394 (54.1)	126 (56.8)	
Former or current	334 (45.9)	96 (43.2)	
Tumor location			<.001
Right upper lobe	247 (33.8)	16 (7.2)	
Right middle lobe	68 (9.3)	1 (0.5)	
Right lower lobe	143 (19.6)	63 (28.4)	
Left upper lobe	148 (20.3)	96 (43.2)	
Left lower lobe	124 (17.0)	46 (20.7)	
Radiologic nodule type			<.001
Part solid	319 (43.7)	148 (66.7)	
Solid	411 (56.3)	74 (33.3)	
Surgical modality			<.001
Sublobar resection	0 (0)	222 (100)	
At least lobectomy	730 (100)	0 (0)	
Tumor histologic type			.02
Adenocarcinoma	644 (88.2)	196 (88.3)	
Squamous cell carcinoma	80 (11.0)	19 (8.6)	
Large cell carcinoma or other histologic types	6 (0.8)	7 (3.2)	
Clinical T category‡			<.001
cT1a	167 (22.9)	104 (46.8)	
cT1b	336 (46.0)	85 (38.3)	
cT1c	227 (31.1)	33 (14.9)	
Solid portion size (mm)†	16 (11–22)	11 (7–16)	<.001
CT to surgery interval (d)†	9 (2–24)	2 (1–24)	.07
Follow-up interval (mo)†	90.4 (72.4–118.9)	85.9 (71.2–107.9)	.03

Note.—Except where indicated, data are numbers of patients, with percentages in parentheses. Data were missing for some variables in the internal set, including history of malignancy other than lung cancer in 0.1% (one of 730 patients), family history of lung cancer in 0.3% (two of 730 patients), and smoking history in 0.3% (two of 730 patients). Patient and tumor characteristics were compared using the Mann-Whitney *U* test, χ^2 test, or Fisher exact test, as appropriate. NA = not available.

* The internal set was divided based on the assignments from the pretraining set as follows: internal training (ie, transfer learning, *n* = 441), internal validation (*n* = 129), and internal testing (*n* = 160).

† Data are medians, with IQRs in parentheses and ranges in brackets.

‡ Clinical T category was determined according to the eighth edition staging system for lung cancer by using the solid portion size on preoperative CT scans.

excluded (2). Sensitivity and specificity were compared using the paired resampling bootstrap procedure (5000 replications). Second, in subgroups of patients meeting criteria for the JCOG trials and CALGB140503 trial, separately, FFR, LCSS, and OS were estimated for DL-based high- and low-risk groups using the Kaplan-Meier method, and the log-rank test was performed between these groups.

The added prognostic value of the DL-driven risk scores to the clinical prognostic factors was investigated using multivariable Cox regression analysis. The hazard ratio of the DL-driven risk score was calculated with adjustment for covariates, such as age, sex, clinical T category, and tumor histologic type, which were determined on the basis of clinical knowledge regarding the prognostic factors in lung cancer. Furthermore, multivariable Cox regression analyses were conducted in a subset of patients with adenocarcinoma, including epidermal growth factor receptor (*EGFR*) mutation status as an extra covariate (Appendix S1).

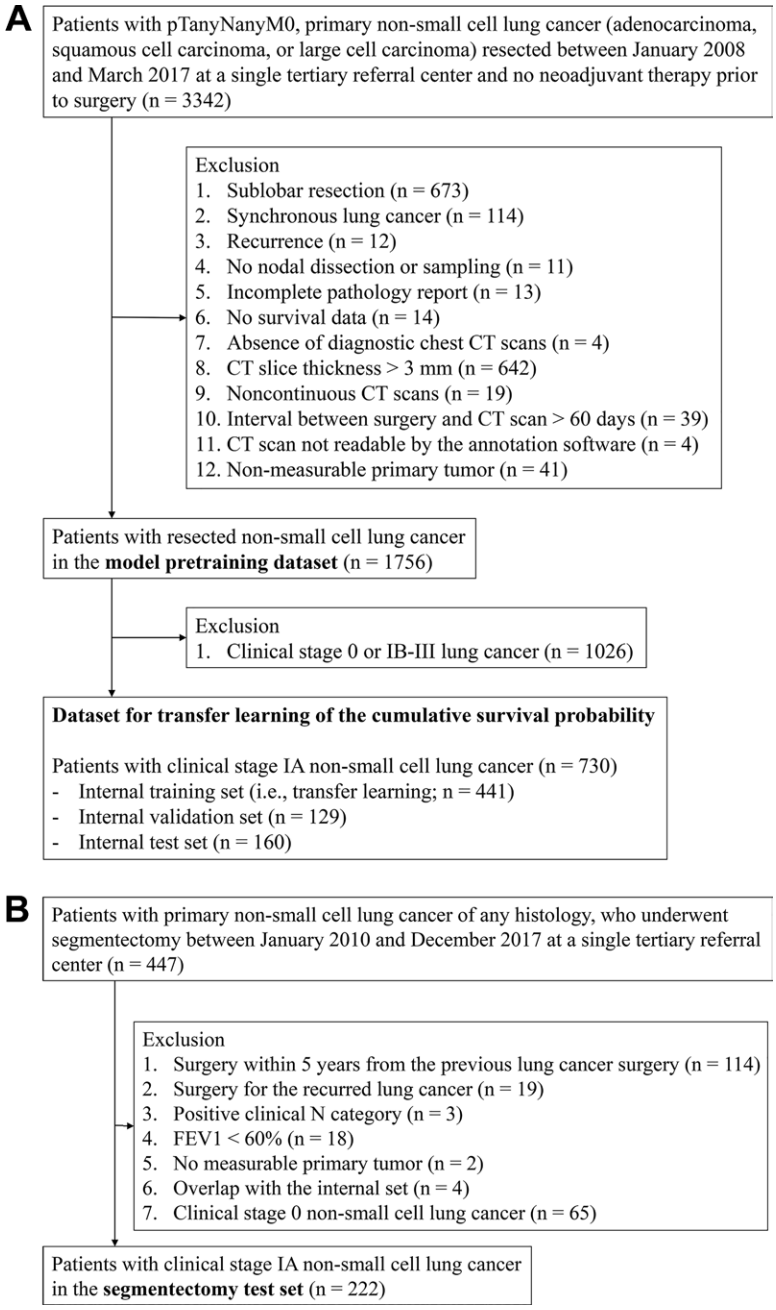

Figure 2: Flowcharts show patient inclusion and exclusion for the **(A)** pretraining and internal set and **(B)** independent segmentectomy test set. The pretraining set **(A)** (including patients with any tumor size and lymph node involvement, but without metastasis, confirmed at pathologic examination [pTanyNanyM0]) was split randomly for training, validation, and testing at a ratio of 6:2:2. Transfer learning was applied to a subset of the pretraining set, specifically in patients with clinical stage IA non–small cell lung cancer (ie, the internal set). The internal set was divided based on the assignments from the pretraining set, such that the ratio of the internal training set (ie, for transfer learning) to the internal validation set to the internal testing set was 6:1.8:2.2. FEV1 = forced expiratory volume in first second of exhalation.

All statistical analyses were performed by one of the authors (H.K.) using R (version 4.3.1; *http://www.R-project.org*) with survival, survminer, and timeROC packages. Sample size calculation was not performed. $P < .05$ was considered indicative of a statistically significance difference. For multiple comparisons, P values were adjusted using the Benjamini-Hochberg procedure.

Results

Study Patients

The pretraining set included 1756 patients with NSCLC (pTanyManyM0) who underwent at least lobectomy (median age, 65 years [IQR, 57–71 years]; 1018 male, 738 female). The internal set for transfer learning included 730 patients with clinical stage IA NSCLC (median age, 63 years [IQR, 56–70 years]; 366 male, 364 female). The independent segmentectomy test set included 222 patients with clinical stage IA NSCLC (median age, 65 years [IQR, 58–71 years]; 114 male, 108 female). The median follow-up duration was 90.4 months (IQR, 72.4–118.9 months) in the internal set and 85.9 months (IQR, 71.2–107.9 months; $P = .03$) in the segmentectomy test set. More characteristics are described in Appendix S2 and Tables 1 and S1. Survival rates are provided in Table S2. Figure 2 shows the patient flowcharts.

Prognostication Using the DL-driven Risk Scores

The median DL-driven risk scores in the independent segmentectomy test set were 0.6% [IQR, 0.1%–2.7%] for 2 years, 2.1% [IQR, 0.5%–8.0%] for 4 years, and 3.9% [IQR, 0.9%–13.4%] for 6 years. The DL model demonstrated robust prognostic discrimination performance across all study outcomes and time points (Table 2, Fig S1). The time-dependent AUC was 0.86 (95% CI: 0.76, 0.96), sensitivity was 87.4% (7.17 of 8.21 patients; 95% CI: 59.4, 100), and specificity was 66.7% (136 of 204 patients; 95% CI: 60.2, 72.8) for 2-year FFR. For 4-year LCSS, the AUC was 0.90 (95% CI: 0.80, 1.00), sensitivity was 100% (5.13 of 5.13 patients; 95% CI: 100, 100), and specificity was 66.5% (137 of 206 patients; 95% CI: 59.6, 73.0). For 4-year OS, the AUC was 0.72 (95% CI: 0.56, 0.88), sensitivity was 71.4% (10 of 14 patients; 95% CI: 45.5, 93.3), and specificity was 65.9% (137 of 208 patients; 95% CI: 59.4, 72.4). The cutoffs were 1.36% for the DL-driven 2-year risk score, 4.36% for the 4-year risk score, and 7.74% for the 6-year risk score. Survival curves are provided in Figures 3 and S2.

Benchmarking of the DL Model against the Randomized Clinical Trial Eligibility Criteria

Among 222 patients in the segmentectomy test set, the DL-defined low-risk group included 141 patients for the DL-driven 2-year, 4-year, and 6-year risk scores, respectively. In addition, 164 patients were eligible for the JCOG trials (ie, 86 patients in JCOG0802, 70 patients in JCOG1211, and eight patients in JCOG0804), and 184 patients were eligible for the CALGB140503 trial.

Table 2: Prognostication Using the DL-driven Risk Scores in the Internal and Independent Segmentectomy Test Sets

Data Set, Outcome, and Time Point	AUC	Sensitivity (%)*	Specificity (%)*	Cutoff (%)
Internal test set (n = 160)†				
Overall survival				
4 years	0.78 (0.63, 0.93)	85.7 (64.7, 100) [12/14]	43.2 (35.4, 51.4) [63/146]	4.36
6 years	0.74 (0.61, 0.87)	80.8 (63.6, 97.9) [17.1/21.2]	44.8 (36.0, 53.6) [56/125]	7.74
Segmentectomy test set (n = 222)				
Freedom from recurrence				
2 years	0.86 (0.76, 0.96)	87.4 (59.4, 100) [7.17/8.21]	66.7 (60.2, 72.8) [136/204]	1.36
4 years	0.85 (0.78, 0.92)	86.4 (65.8, 100) [13.5/15.6]	69.0 (62.0, 75.3) [129/187]	4.36
Lung cancer–specific survival				
4 years	0.90 (0.80, 1.00)	100 (100, 100) [5.13/5.13]	66.5 (59.6, 73.0) [137/206]	4.36
6 years	0.86 (0.78, 0.95)	89.9 (66.4, 100) [9.42/10.5]	67.3 (60.0, 74.3) [109/162]	7.74
Overall survival				
4 years	0.72 (0.56, 0.88)	71.4 (45.5, 93.3) [10/14]	65.9 (59.4, 72.4) [137/208]	4.36
6 years	0.78 (0.68, 0.88)	76.1 (59.3, 91.0) [22.3/29.3]	67.3 (60.0, 74.3) [109/162]	7.74

Note.—Data in parentheses are 95% CIs. AUC, sensitivity, and specificity are time-dependent measures for the event of interest (ie, outcome). The median values of each risk score in the internal validation set were used as cutoffs to calculate sensitivity and specificity. The median DL-driven risk scores for the internal test set were 2.0% (IQR, 0.5%–5.3%) for 2 years, 6.0% (IQR, 1.8%–13.6%) for 4 years, and 10.4% (IQR, 3.4%–22.0%) for 6 years. The median DL-driven risk scores for the independent segmentectomy test set were 0.6% (IQR, 0.1%–2.7%) for 2 years, 2.1% (IQR, 0.5%–8.0%) for 4 years, and 3.9% (IQR, 0.9%–13.4%) for 6 years. AUC = area under the receiver operating characteristic curve, DL = deep learning.

* Except where indicated, data are percentages, with estimated numbers of patients in brackets calculated using the inverse probability of censoring weighting approach.

† The internal set (n = 730) was divided based on the assignments from the pretraining set as follows: internal training (ie, transfer learning, n = 441), internal validation (n = 129), and internal testing (n = 160).

The DL model showed higher time-dependent sensitivity for FFR, LCSS, and OS than the JCOG criteria (2-year FFR, 87.4% vs 37.6% [3.08 of 8.21 patients; 95% CI: 0, 74.5; P = .02]; 4-year LCSS, 100% vs 39.9% [2.05 of 5.13 patients; 95% CI: 0, 100; P < .001]; 4-year OS, 71.4% vs 28.6% [four of 14 patients; 95% CI: 6.70, 54.6; P = .005]) at similar specificity (2-year FFR, 66.7% vs 75.5% [154 of 204 patients; 95% CI: 68.8, 81.0; P = .05]; 4-year LCSS, 66.5% vs 74.3% [153 of 206 patients; 95% CI: 68.3, 80.0; P = .05]; 4-year OS, 65.9% vs 74.0% [154 of 208 patients; 95% CI: 67.9, 79.8; P = .05]) (Table 3).

Similar results were noted when comparing the sensitivity of the DL model with that of the CALGB140503 criterion (2-year FFR, 87.4% vs 37.6% [3.08 of 8.21 patients; 95% CI: 0, 74.5; P = .02]; 4-year LCSS, 100% vs 39.9% [2.05 of 5.13 patients; 95% CI: 0, 100; P < .001]; 4-year OS, 71.4% vs 21.4% [three of 14 patients; 95% CI: 0, 45.8; P < .001]) (Table 3). However, the specificity of the DL model was lower than that of the CALGB140503 criterion for the study outcomes (2-year FFR, 66.7% vs 84.8% [173 of 204 patients; 95% CI: 80.0, 89.2; P < .001]; 4-year LCSS, 66.5% vs 83.5% [172 of 206 patients; 95% CI: 77.9, 88.2; P < .001]; 4-year OS, 65.9% vs 83.2% [173 of 208 patients; 95% CI: 78.0, 88.1; P < .001]).

Subgroup Analyses in the Randomized Clinical Trial–eligible Patients

Kaplan-Meier survival curves stratified according to the DL-based risk groups are provided in Figure 4. The DL-based low-risk groups, stratified using the DL-driven 2-year, 4-year, and 6-year risk scores, showed better outcomes than the DL-based high-risk groups for FFR, LCSS, and OS in patients eligible for the CALGB140503 trial (n = 184, all P < .001) and patients eligible for the JCOG trials (n = 164, all P < .001) (Table S3).

Multivariable Cox Regression Analyses

The multivariable Cox regression analyses revealed that the DL risk scores remained significant prognostic factors in the segmentectomy test set for FFR, LCSS, and OS after adjusting for other clinical factors (Tables 4, S4). The adjusted hazard ratios of the DL risk score (per percentage) were 1.13 (95% CI: 1.05, 1.23; P = .002) for FFR, 1.09 (95% CI: 1.03, 1.14; P = .001) for LCSS, and 1.06 (95% CI: 1.03, 1.09; P < .001) for OS. Representative CT images with heat maps are provided in Figure 5.

Multivariable Cox Regression Analyses in Patients with Adenocarcinoma

The prognostic value of the DL risk score remained significant after adjusting for *EGFR* mutation status (adjusted hazard ratio, 1.32 [95% CI: 1.18, 1.49; P < .001] for FFR; 1.16 [95% CI: 1.09, 1.23; P < .001] for LCSS; 1.07 [95% CI: 1.03, 1.11; P = .001] for OS) (Appendix S2; Tables S5, S6).

Discussion

Recent clinical trials have shown that for small peripheral non–small cell lung cancer (NSCLC) tumors, sublobar resection might be as effective as lobectomy, with similar survival rates

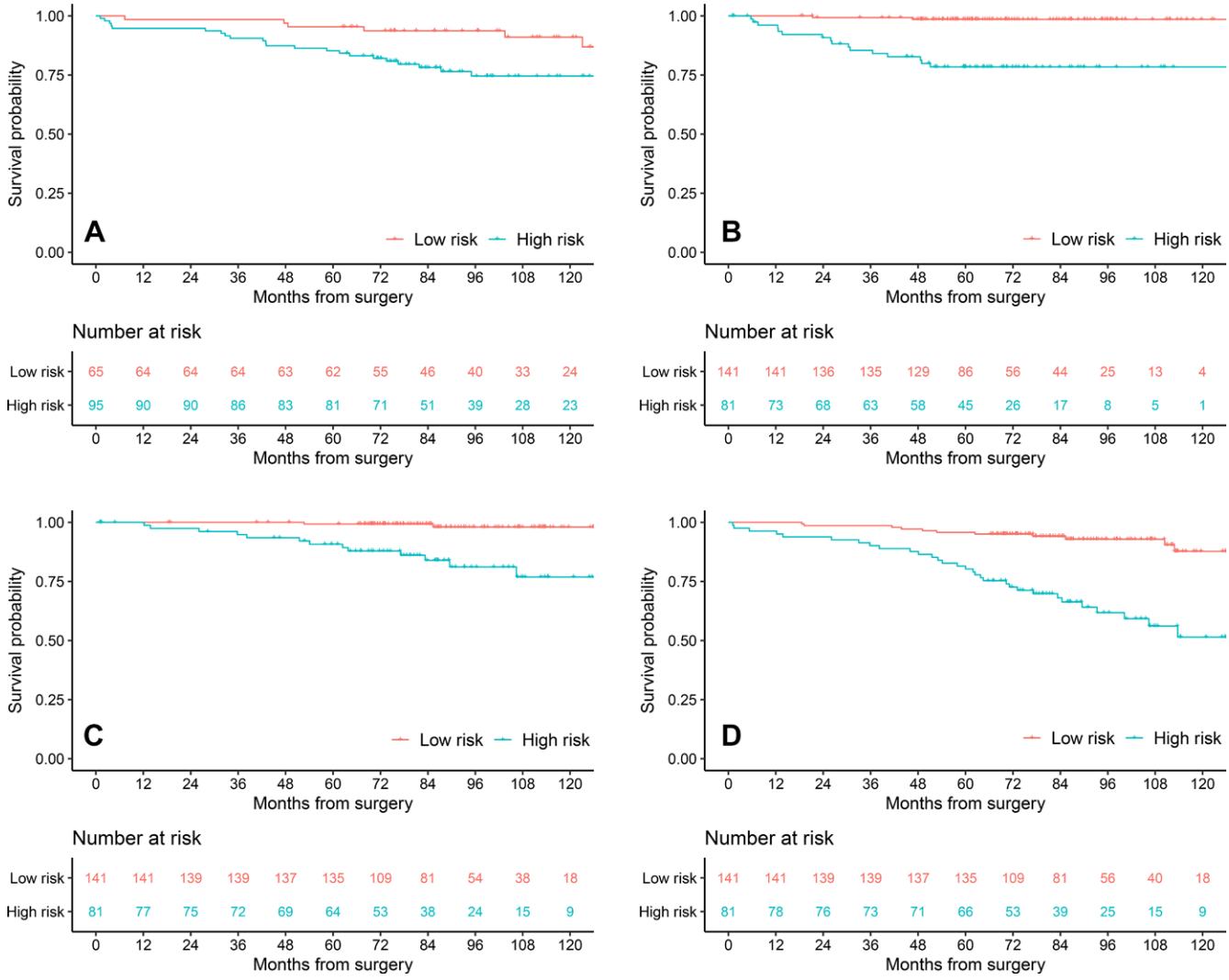

Figure 3: Kaplan-Meier survival curves stratified according to the dichotomized deep learning (DL)–driven risk scores show **(A)** overall survival (OS) in the internal test set using the DL-driven 4-year risk scores and **(B)** freedom from recurrence, **(C)** lung cancer–specific survival, and **(D)** OS in the segmentectomy test set using the DL-driven 2-year, 4-year, and 6-year risk scores. The cutoffs were determined empirically as the median values in the internal validation set, which were 1.36% for the DL-driven 2-year risk score and 4.36% for the 4-year risk score. The cutoffs remained unchanged regardless of the study outcome.

(2,3). However, further risk stratification in patients undergoing segmentectomy is currently limited. In this retrospective study, we developed and validated a deep learning (DL) model with preoperative CT scans and four clinicoradiologic variables (age, sex, solid portion size, and consolidation to tumor ratio) as inputs for risk stratification in 222 patients with clinical stage IA NSCLC undergoing segmentectomy. The model demonstrated robust prognostication performance across three oncologic outcomes, including freedom from recurrence, lung cancer–specific survival, and overall survival. For recurrence within 2 years, the time-dependent sensitivity was 87.4% and specificity was 66.7%, with an area under the receiver operating characteristic curve of 0.86. Therefore, our model detected most of the individuals at high risk who were vulnerable to disease recurrence. In this respect, the sensitivity and specificity of the model for recurrence or death were compared with those of the clinical trial eligibility criteria for sublobar resection. DL-based risk stratification demonstrated higher sensitivity for 2-year recurrence than the Japan Clinical Oncology Group eligibility criteria (87.4% vs 37.6%, $P = .02$), while maintaining similar specificity. In addition, in subgroups of clinical trial–eligible patients, the DL model was able to further stratify those at high and low risk.

In clinical practice, anatomic information on preoperative CT scans is qualitatively assessed by radiologists, pulmonologists, and surgeons for the purpose of clinical staging. However, CT scans contain more data beyond tumor dimension, density, and location. A few recent studies have shown that prognostic information can be extracted quantitatively (14,15,29–31). Considering that CT scans are universally acquired for staging in patients with lung cancer, fully exploiting all available data through a model-based approach is reasonable. In fact, risk stratification in lung cancer by simple dimensional measurements of tumors may not be an optimal strategy. Although 2 cm was used as a diameter threshold in both the JCOG0802 and CALGB140503 trials, there is some evidence that segmentectomy may be feasible in tumors larger than 2 cm (ie, 2–3

Table 3: Risk Stratification Benchmarking Against the Randomized Clinical Trial Eligibility Criteria

Outcome and Time Point	JCOG Trials*				CALGB140503†			
	Sensitivity	*P* Value	Specificity	*P* Value	Sensitivity	*P* Value	Specificity	*P* Value
Freedom from recurrence								
2 years	37.6 (0, 74.5) [3.08/8.21]	.02	75.5 (68.8, 81.0) [154/204]	.05	37.6 (0, 74.5) [3.08/8.21]	.02	84.8 (80.0, 89.2) [173/204]	<.001
4 years	47.1 (20.4, 73.7) [7.35/15.6]	.03	76.5 (70.4, 83.0) [143/187]	.05	47.1 (20.4, 73.7) [7.35/15.6]	.03	86.1 (80.8, 91.1) [161/187]	<.001
Lung cancer–specific survival								
4 years	39.9 (0, 100) [2.05/5.13]	<.001	74.3 (68.3, 80.0) [153/206]	.05	39.9 (0, 100) [2.05/5.13]	<.001	83.5 (77.9, 88.2) [172/206]	<.001
6 years	50.4 (17.1, 83.4) [5.28/10.5]	.02	75.3 (68.8, 81.8) [122/162]	.05	50.4 (17.1, 83.4) [5.28/10.5]	.02	84.6 (78.9, 89.8) [137/162]	<.001
Overall survival								
4 years	28.6 (6.70, 54.6) [4/14]	.005	74.0 (67.9, 79.8) [154/208]	.05	21.4 (0, 45.8) [3/14]	.001	83.2 (78.0, 88.1) [173/208]	<.001
6 years	41.9 (24.4, 61.1) [12.3/29.3]	<.001	75.3 (68.7, 81.7) [122/162]	.05	38.5 (21.5, 56.9) [11.3/29.3]	<.001	84.6 (78.9, 89.9) [137/162]	<.001

Note.—Except where indicated, data are percentages, with 95% CIs in parentheses and estimated numbers of patients in brackets calculated using the inverse probability of censoring weighting approach. Time-dependent sensitivity and specificity were calculated for each outcome. *P* values indicate comparisons of the diagnostic performance measures between the clinical trial eligibility criteria and DL-based risk groups using a paired resampling bootstrap procedure (5000 replications), which accounts for the paired nature of the data by resampling matched pairs of observations. *P* values were adjusted for multiple comparisons across time and different outcomes. Sensitivity and specificity of the DL model are provided in Table 2. Note that the sample size for calculating sensitivity in the present study was insufficient to yield reliable CIs. CALGB = Cancer and Leukemia Group B, DL = deep learning, JCOG = Japan Clinical Oncology Group.

* Eligibility criteria in the JCOG trials are a composite, encompassing the criteria from the JCOG0802, JCOG1211, and JCOG0804 trials (tumors ≤3 cm, except for those >2 cm with a consolidation to tumor ratio >0.5).

† CALGB140503 trial eligibility included tumors with a solid component measuring 2 cm or less.

cm) (32,33). A DL model that analyzes all pixels in and around the tumor may enable the identification of individuals at high risk among those with tumors smaller than 2 cm and individuals at low risk among those with tumors larger than 2 cm. Furthermore, the DL risk scores remained significant for predicting clinical outcomes after adjusting for the status of sensitizing *EGFR* mutations in patients with adenocarcinoma, indicating the utility of the model in patients with and without the oncogenic driver mutations.

Our study had several strengths. First, the DL model was pretrained to predict cumulative survival in a large number of patients with NSCLC, which was supported by histopathologic feature learning. The selection of the histopathologic labels was based on evidence from published research, including the results from the JCOG0201 studies (17–22). Second, the transfer learning used the solid portion size and consolidation to tumor ratio, following evidence from recent clinical trials (2,3,23). Third, we analyzed three study outcomes and conducted separate evaluations of recurrence and mortality, in accordance with the strategy used in the clinical trial (3).

Nonetheless, our study had limitations. First, the DL model was developed and validated using data sets from the same hospital. Although there was no overlap between the internal and segmentectomy test sets, true external validation (using an external test set) was not carried out. Second, survival analysis comparing segmentectomy and lobectomy was not conducted for the high- and low-risk groups identified by the DL model. Therefore, it is not feasible to assert that the DL model is suitable for selecting candidates for segmentectomy. Third, the DL model was developed using the lobectomy cohort and validated in the segmentectomy test set. This approach assumed that there are common risk factors in patients with early-stage lung cancer independent from the surgical modality. A similar strategy was used in the JCOG0802, JCOG1211, and JCOG0804 trials, following the results of JCOG0201, which was a prospective study in patients treated with at least lobectomy. Fourth, the DL model was not fully automatic. Finally, eligibility for segmentectomy is not determined solely based on tumor dimensions. The tumor location in the periphery and the absence of nodal metastasis are essential prerequisites, and these factors were not evaluated in our study.

In conclusion, the CT-based deep learning model identified patients at high risk among those with clinical stage IA non–small cell lung cancer who underwent segmentectomy, demonstrating higher sensitivity for recurrence and/or death compared with clinical trial eligibility criteria for sublobar resection.

Acknowledgments: We thank Myunghee Lee, BA, for invaluable assistance in data acquisition, and Yunhee Choi, PhD, for expert consultation on statistical analysis. During the revision of the manuscript, the English grammar was checked and proofread by ChatGPT (GPT-4; OpenAI).

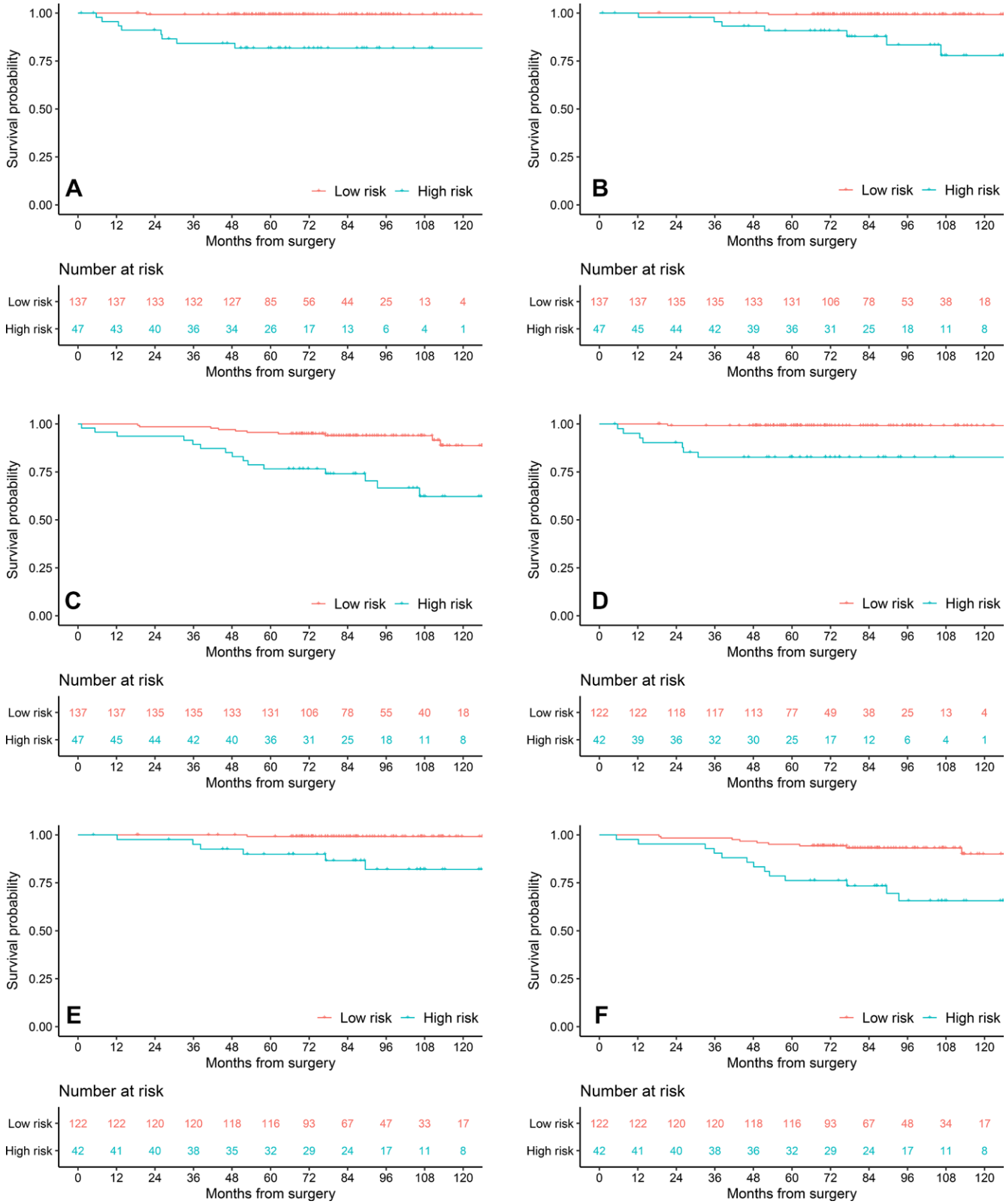

Figure 4: Kaplan-Meier survival curves according to the dichotomized deep learning (DL)–driven risk scores in segmentectomy subgroups of patients who met clinical trial eligibility. **(A–C)** Graphs show freedom from recurrence (FFR) **(A)**, lung cancer–specific survival (LCSS) **(B)**, and overall survival (OS) **(C)** in patients eligible for the Cancer and Leukemia Group B 140503 trial. **(D–F)** Graphs show FFR **(D)**, LCSS **(E)**, and OS **(F)** in patients eligible for the Japan Clinical Oncology Group (JCOG) trials (JCOG0802, JCOG1211, and JCOG0804). The DL-driven 2-year risk score was used for FFR, and the 4-year risk score was used for LCSS and OS. The cutoffs were determined empirically as the median values in the internal validation set, which were 1.36% for the DL-driven 2-year risk score and 4.36% for the 4-year risk score. The cutoffs were not altered according to the study outcomes. In the segmentectomy test set, the same patients were consistently classified into the DL-based low-risk group across different time points (2, 4, and 6 years).

Table 4: Multivariable Cox Regression Analyses for the Independent Segmentectomy Test Set

	Recurrence		Lung Cancer–specific Mortality		Overall Mortality	
Variable	Hazard Ratio	*P* Value	Hazard Ratio	*P* Value	Hazard Ratio	*P* Value
Age (per year)	1.09 (1.02, 1.16)	.01	1.12 (1.04, 1.21)	.003	1.08 (1.03, 1.13)	<.001
Male sex (reference: female)	1.01 (0.29, 3.50)	.99	0.96 (0.24, 3.89)	.95	2.02 (0.84, 4.84)	.11
Clinical T1c (reference: T1a or T1b)	3.18 (1.13, 8.93)	.03	2.16 (0.72, 6.49)	.17	1.76 (0.88, 3.52)	.11
Adenocarcinoma (reference: nonadenocarcinoma)	5.12 (1.12, 23.3)	.03	4.28 (0.92, 19.9)	.06	1.84 (0.80, 4.25)	.15
DL risk score (per percentage point)*	1.13 (1.05, 1.23)	.002	1.09 (1.03, 1.14)	.001	1.06 (1.03, 1.09)	<.001

Note.—Data in parentheses are 95% CIs. DL = deep learning.

* Multivariable Cox regression analyses were performed using the DL-driven 2-year risk score for freedom from recurrence and DL-driven 4-year risk score for lung cancer–specific survival and overall survival.

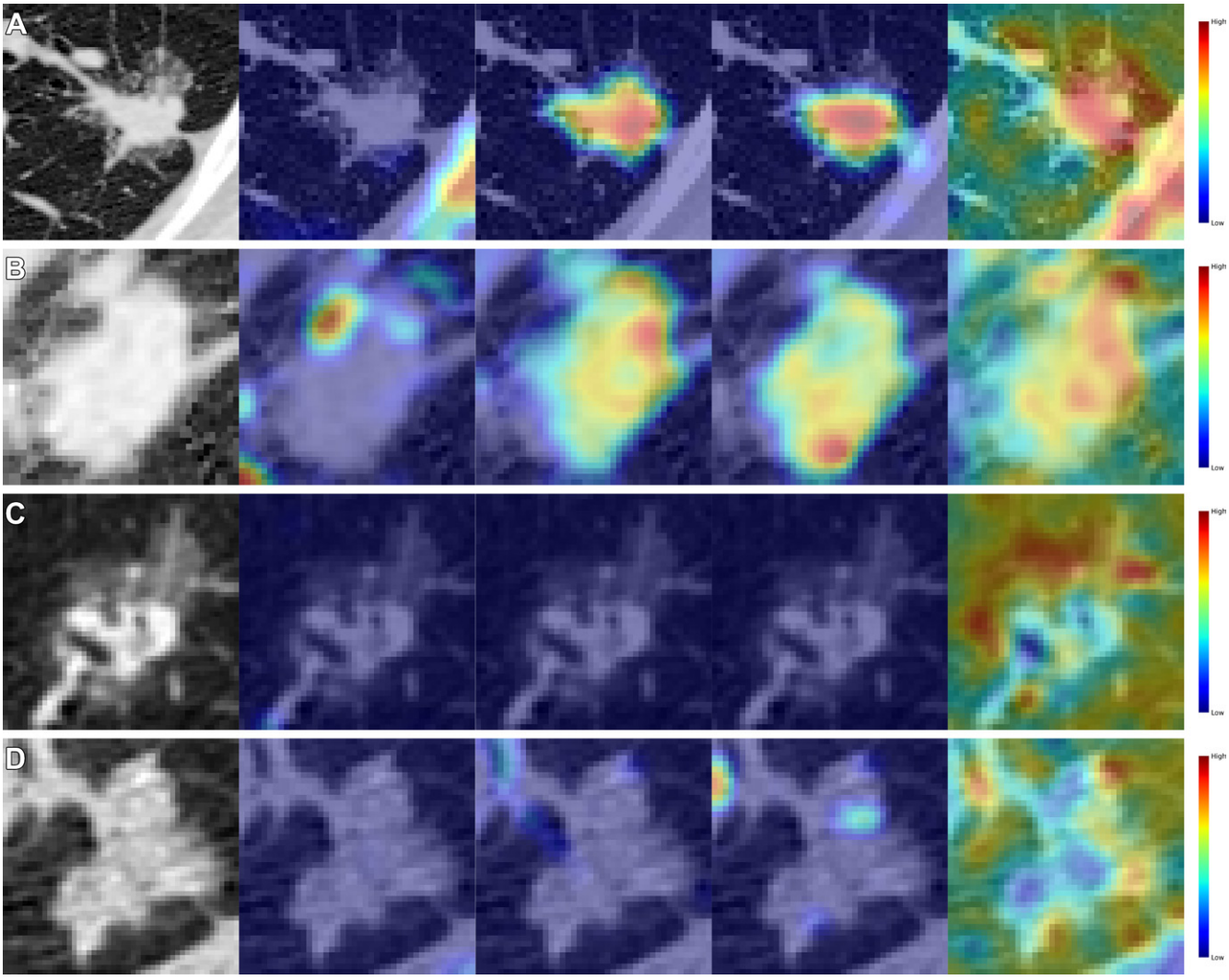

Figure 5: Representative CT images with heat map visualization. From left to right: Axial nonenhanced CT images show a preoperative scan with overlaid gradient-weighted activation maps for visceral pleural invasion, lymphovascular invasion, lymph node, and survival prediction, respectively. **(A)** Images in an 83-year-old male patient with clinical stage IA3 adenocarcinoma. The deep learning (DL)–driven 2-year risk score was 3.65% and the 4-year risk score was 10.2%. The tumor recurred 36.5 months after surgery. **(B)** Images in a 79-year-old female patient with clinical stage IA2 adenocarcinoma. The DL-driven 2-year risk score was 8.60% and the 4-year risk score was 20.2%. Tumor recurrence was observed 25.9 months after surgery. **(C)** Images in a 71-year-old male patient with clinical stage IA2 adenocarcinoma. The DL-driven 2-year risk score was 0.16% and the 4-year risk score was 0.74%. There was no evidence of disease recurrence until 60.8 months of postoperative follow-up. **(D)** Images in a 76-year-old female patient with clinical stage IA3 adenocarcinoma. The DL-driven 2-year risk score was 0.63% and the 4-year risk score was 2.28%. No recurrence was noted at a 22-month follow-up visit. The DL model predicted the cumulative overall survival probability in patients with clinical stage IA lung cancer, and the prediction was enhanced by the multitask learning of CT features for visceral pleural invasion, lymphovascular invasion, and lymph node metastasis. The color bar transitions from dark blue to dark red, indicating pixel activation ranging from a low to high degree on the heat maps.

Author contributions: Guarantor of integrity of entire study, **H.K.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **J.M.G., H.K.**; clinical studies, all authors; experimental studies, **H.K.**; statistical analysis, **K.J.N., H.K.**; and manuscript editing, all authors

Disclosures of conflicts of interest: K.J.N. Cofounder and chief medical officer of Portrai. **Y.T.K.** Consulting fees from Johnson & Johnson; lecture payment from AstraZeneca; holds stock option in Genome Insight. **J.M.G.** Research grants from LG Electronics and Coreline Soft; associate editor of thoracic imaging for *Radiology*. **H.K.** Research grants from RADISEN and Kakao Brain; consulting fees from RADISEN; holds stock and stock option in MEDICALIP.

References

1. Ginsberg RJ, Rubinstein LV. Randomized trial of lobectomy versus limited resection for T1 N0 non-small cell lung cancer. Lung Cancer Study Group. Ann Thorac Surg 1995;60(3):615–622; discussion 622–623.
2. Altorki N, Wang X, Kozono D, et al. Lobar or sublobar resection for peripheral stage IA non-small-cell lung cancer. N Engl J Med 2023;388(6):489–498.
3. Saji H, Okada M, Tsuboi M, et al. Segmentectomy versus lobectomy in small-sized peripheral non-small-cell lung cancer (JCOG0802/WJOG4607L): a multicentre, open-label, phase 3, randomised, controlled, non-inferiority trial. Lancet 2022;399(10335):1607–1617.
4. Ten Haaf K, Jeon J, Tammemägi MC, et al. Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study. PLoS Med 2017;14(4):e1002277. [Published correction appears in PLoS Med 2020;17(9):e1003403.]
5. Callister ME, Baldwin DR, Akram AR, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. Thorax 2015;70(Suppl 2):ii1–ii54. [Published correction appears in Thorax 2015;70(12):1188.]
6. Gould MK, Donington J, Lynch WR, et al. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 2013;143(5 Suppl):e93S–e120S.
7. Lee JH, Lee D, Lu MT, et al. Deep learning to optimize candidate selection for lung cancer CT screening: advancing the 2021 USPSTF recommendations. Radiology 2022;305(1):209–218.
8. Goldstraw P, Chansky K, Crowley J, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol 2016;11(1):39–51.
9. Choi H, Kim H, Park CM, Kim YT, Goo JM. Central tumor location at chest CT is an adverse prognostic factor for disease-free survival of node-negative early-stage lung adenocarcinomas. Radiology 2021;299(2):438–447.
10. Kim H, Choi H, Lee KH, et al. Definitions of central tumors in radiologically node-negative, early-stage lung cancer for preoperative mediastinal lymph node staging: a dual-institution, multireader study. Chest 2022;161(5):1393–1406.
11. Kim H, Goo JM, Kim YT, Park CM. Clinical T category of non-small cell lung cancers: prognostic performance of unidimensional versus bidimensional measurements at CT. Radiology 2019;290(3):807–813.
12. Kim H, Goo JM, Kim YT, Park CM. CT-defined visceral pleural invasion in T1 lung adenocarcinoma: lack of relationship to disease-free survival. Radiology 2019;292(3):741–749.
13. Kim H, Goo JM, Kim YT, Park CM. Validation of the Eighth Edition Clinical T Categorization System for Clinical Stage IA, Resected Lung Adenocarcinomas: Prognostic Implications of the Ground-Glass Opacity Component. J Thorac Oncol 2020;15(4):580–588.
14. Kim H, Goo JM, Lee KH, Kim YT, Park CM. Preoperative CT-based deep learning model for predicting disease-free survival in patients with lung adenocarcinomas. Radiology 2020;296(1):216–224.
15. Lee KH, Lee JH, Park S, et al. Computed tomography-based prognostication in lung adenocarcinomas through histopathological feature learning: a retrospective multicenter study. Ann Am Thorac Soc 2023;20(7):1020–1028.
16. Nam JG, Hong H, Choi SH, et al. No prognostic impact of staging brain MRI in patients with stage IA non-small cell lung cancer. Radiology 2022;303(3):632–643.
17. Asamura H, Hishida T, Suzuki K, et al. Radiographically determined noninvasive adenocarcinoma of the lung: survival outcomes of Japan Clinical Oncology Group 0201. J Thorac Cardiovasc Surg 2013;146(1):24–30.
18. Higgins KA, Chino JP, Ready N, et al. Lymphovascular invasion in non-small-cell lung cancer: implications for staging and adjuvant therapy. J Thorac Oncol 2012;7(7):1141–1147.
19. Kawase A, Yoshida J, Miyaoka E, et al. Visceral pleural invasion classification in non-small-cell lung cancer in the 7th edition of the tumor, node, metastasis classification for lung cancer: validation analysis based on a large-scale nationwide database. J Thorac Oncol 2013;8(5):606–611.
20. Pathak R, Goldberg SB, Canavan M, et al. Association of survival with adjuvant chemotherapy among patients with early-stage non-small cell lung cancer with vs without high-risk clinicopathologic features. JAMA Oncol 2020;6(11):1741–1750.
21. Suzuki K, Koike T, Asakawa T, et al. A prospective radiological study of thin-section computed tomography to predict pathological noninvasiveness in peripheral clinical IA lung cancer (Japan Clinical Oncology Group 0201). J Thorac Oncol 2011;6(4):751–756.
22. Travis WD, Brambilla E, Rami-Porta R, et al. Visceral pleural invasion: pathologic criteria and use of elastic stains: proposal for the 7th edition of the TNM classification for lung cancer. J Thorac Oncol 2008;3(12):1384–1390.
23. Aokage K, Suzuki K, Saji H, et al. Segmentectomy for ground-glass-dominant lung cancer with a tumour diameter of 3 cm or less including ground-glass opacity (JCOG1211): a multicentre, single-arm, confirmatory, phase 3 trial. Lancet Respir Med 2023;11(6):540–549.
24. Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. PeerJ 2019;7:e6257.
25. Hung JJ, Yeh YC, Jeng WJ, et al. Predictive value of the international association for the study of lung cancer/American Thoracic Society/European Respiratory Society classification of lung adenocarcinoma in tumor recurrence and patient survival. J Clin Oncol 2014;32(22):2357–2364.
26. Punt CJ, Buyse M, Köhne CH, et al. Endpoints in adjuvant treatment trials: a systematic review of the literature in colon cancer and proposed definitions for future trials. J Natl Cancer Inst 2007;99(13):998–1003.
27. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med 2013;32(30):5381–5397.
28. Suzuki K, Watanabe SI, Wakabayashi M, et al. A single-arm study of sublobar resection for ground-glass opacity dominant peripheral lung cancer. J Thorac Cardiovasc Surg 2022;163(1):289–301.e2.
29. Chen W, Hou X, Hu Y, Huang G, Ye X, Nie S. A deep learning- and CT image-based prognostic model for the prediction of survival in non-small cell lung cancer. Med Phys 2021;48(12):7946–7958.
30. Huang B, Sollee J, Luo YH, et al. Prediction of lung malignancy progression and survival with machine learning based on pre-treatment FDG-PET/CT. EBioMedicine 2022;82:104127.
31. Zheng S, Guo J, Langendijk JA, et al. Survival prediction for stage I-IIIA non-small cell lung cancer using deep learning. Radiother Oncol 2023;180:109483.
32. Chan EG, Chan PG, Mazur SN, et al. Outcomes with segmentectomy versus lobectomy in patients with clinical T1cN0M0 non-small cell lung cancer. J Thorac Cardiovasc Surg 2021;161(5):1639–1648.e2.
33. Hattori A, Matsunaga T, Fukui M, Takamochi K, Oh S, Suzuki K. Oncologic outcomes of segmentectomy for stage IA radiological solid-predominant lung cancer >2 cm in maximum tumour size. Interact Cardiovasc Thorac Surg 2022;35(6):ivac246.