Deep-Learning Model for Real-Time Prediction of Recurrence in Early-Stage Non–Small Cell Lung Cancer: A Multimodal Approach (RADAR CARE Study)

Hyun Ae Jung, MD, PhD[1]; Daehwan Lee, BS[2]; Boram Park, PhD[3,4]; Kiwon Lee, MS[2]; Ho Yun Lee, MD, PhD[5]; Tae Jung Kim, MD, PhD[5]; Yeong Jeong Jeon, MD, PhD[6]; Junghee Lee, MD, PhD[6]; Seong Yong Park, MD, PhD[6]; Jong Ho Cho, MD, PhD[6]; Yong Soo Choi, MD, PhD[6]; Sehhoon Park, MD, PhD[1]; Jong-Mu Sun, MD, PhD[1]; Se-Hoon Lee, MD, PhD[1]; Jin Seok Ahn, MD, PhD[1]; Myung-Ju Ahn, MD, PhD[1]; and Hong Kwan Kim, MD, PhD[5]

DOI https://doi.org/10.1200/PO-25-00172

ABSTRACT

PURPOSE The surveillance protocol for early-stage non–small cell lung cancer (NSCLC) is not contingent upon individualized risk factors for recurrence. This study aimed to use comprehensive data from clinical practice to develop a deep-learning model for practical longitudinal monitoring.

METHODS A multimodal deep-learning model with transformers was developed for real-time recurrence prediction using baseline clinical, pathological, and molecular data with longitudinal laboratory and radiologic data collected during surveillance. Patients with NSCLC (stage I to III) who underwent surgery with curative intent between January 2008 and September 2022 were included. The primary outcome was predicting recurrence within 1 year after the monitoring point. This study demonstrates the timely provision of risk scores (RADAR score) and determined thresholds and the corresponding AUC.

RESULTS A total of 14,177 patients were enrolled (10,262 with stage I, 2,380 with stage II, and 1,703 with stage III). The model incorporated 64 clinical-pathological-molecular factors at baseline, along with longitudinal laboratory and computed tomography imaging interpretation data. The mean baseline RADAR score was 0.324 (standard deviation [SD], 0.256) in stage I, 0.660 (SD, 0.210) in stage II, and 0.824 (SD, 0.140) in stage III. The AUC for predicting relapse within 1 year of the monitoring point was 0.854 across all stages, with a sensitivity of 86.0% and a specificity of 71.3% (AUC = 0.872 in stage I, AUC = 0.737 in stage II, and AUC = 0.724 in stage III).

CONCLUSION This pilot study introduces a deep-learning model that uses multimodal data from routine clinical practice to predict relapses in early-stage NSCLC. It demonstrates the timely provision of RADAR risk scores to clinicians for recurrence prediction, potentially guiding risk-adapted surveillance strategies and aggressive adjuvant systemic treatment.

ACCOMPANYING CONTENT

Data Supplement

Visual Abstract

Accepted June 10, 2025
Published July 23, 2025

JCO Precis Oncol 9:e2500172

INTRODUCTION

In Republic of Korea, 28,949 cases of lung cancer occur per 100,000 people, of which non–small cell lung cancer (NSCLC) accounts for 84.1%, and early lung cancer accounts for approximately 34.6%.[1] Surgical resection with curative intent with or without adjuvant chemotherapy remains the standard of care for early-stage NSCLC. The TNM stage is the most commonly used to predict the prognosis of NSCLC after surgical resection with curative intent, with the adjuvant treatment strategy including postoperative adjuvant chemotherapy or radiotherapy determined on the basis of the pathologic stage.[2] However, NSCLC ranks first in cancer-related mortality, with a 5-year survival rate of 36.8%, which is poor compared with other solid tumors, such as 5-year survival rates of 93.8% in breast cancer, 74.3% in colon cancer, and 77.9% in stomach cancer. The 5-year survival rates for stages 1A1, 1A2, 1A3, 1B, 2A, 2B, and 3A are 90%, 85%, 80%, 73%, 65%, 56%, and 41%, respectively.[3]

Differences in tumor biology, such as *EGFR*, *ALK*, *KRAS*, *MET*, and *TP53* status; and prognosis, can vary within the same stage.[4] Moreover, aggressive adjuvant treatment with targeted therapy or immunotherapy has shown efficacy in

CONTEXT

Key Objective

Can a transformer-based deep-learning model, leveraging multimodal real-world clinical data, predict recurrence within 1 year in real time for patients with early-stage non–small cell lung cancer (NSCLC) following curative surgery?

Knowledge Generated

The RADAR CARE study developed and validated a recurrence prediction model using data from 14,177 patients, incorporating baseline clinical, pathological, and genomic information, along with longitudinal laboratory and radiologic data. The model demonstrated strong predictive performance (AUC, 0.854) and revealed four distinct RADAR score trajectories that may inform surveillance strategies and guide decisions on adjuvant therapy. Importantly, the RADAR score predicted recurrence independent of TNM stage.

Relevance

This study highlights the potential for artificial intelligence–powered tools to be integrated into routine oncology practice, enabling personalized recurrence risk assessment in early-stage NSCLC. The RADAR score may support clinicians in adjusting surveillance frequency and identifying patients who may benefit from early, risk-adapted adjuvant treatment.

preventing disease recurrence. Adjuvant osimertinib for 3 years in stage I to IIIA *EGFR* mutation–positive NSCLC reduced the risk of recurrence by 73% and showed superior disease-free survival (DFS) compared with placebo (median DFS 65.8 months for osimertinib *v* 28.1 months for placebo) in the ADAURA study.[5] Adjuvant 3-year alectinib treatment reduced the risk of recurrence by 76% in *ALK*-positive NSCLC and showed DFS superior to that of placebo in the ALINA study.[6] Adjuvant atezolizumab reduced the risk of recurrence by 34% in cases of NSCLC with >1% PD-L1 expression.[7] A practical model for longitudinal monitoring for recurrence is needed to enable timely, risk-adapted treatment such as targeted therapy or immunotherapy while balancing the potential trade-off between longer treatment duration and reduced risk of recurrence on one side with the increased burden of treatment and side effects on the other.

Artificial intelligence (AI) has shown great promise for important applications in medicine. In lung cancer research, several studies have shown that AI is useful for radiologic or pathologic diagnosis and interpretation and for predicting the therapeutic efficacy of immunotherapy.[8] The Sybil study reported that a deep-learning model could use chest computed tomography (CT) scans to predict the occurrence of lung cancer 1-6 years after a screening.[9]

This study used a multimodal approach to develop a deep-learning model for real-time prediction of recurrence in early-stage NSCLC.

METHODS

Study Participants and Data Collection

Real-time autOmatically updated data warehOuse for healThcare in Head and neck cancer, Esophageal cancer, Lung cancer, Thymic cancer and Mesothelioma (ROOT-HEALTH) was developed to extract and update clinical data from a clinical data warehouse. It was successfully developed and validated through single- and multicenter studies.[10-12] This study collected ROOT-HEALTH clinicopathologic genomic data for patients with pathologic stage I to IIIA NSCLC who underwent surgical resection with curative intent at Samsung Medical Center between January 2008 and September 2022. We gathered baseline data before and at the time of surgery and longitudinal laboratory and CT imaging interpretation data during surveillance.

Study Scheme for the RADAR Study

Figure 1A shows the scheme for the RADAR CARE (Real-time Risk-Adapted Surveillance Comprehensive Strategy AI Model for Early-Stage NSCLC) study to develop a risk model for recurrence of early-stage NSCLC using both baseline data from the time of surgery and longitudinal data collected during follow-up. The current surveillance practice for NSCLC is regular radiologic monitoring every 3-6 months without predicting recurrence or implementing an early intervention strategy. The RADAR CARE model predicts the risk of recurrence 1 year after the monitoring point, and the RADAR score can be used to guide monitoring intervals for early detection and intervention.

Data Preprocessing and Labeling

Several preprocessing steps were used to efficiently process clinical and pathological data, laboratory data, and CT imaging interpretation data. The clinical and pathological data are categorical and were transformed into 130-dimensional vectors using one-hot encoding. The laboratory data are numerical and were subjected to outlier treatment, normalized to values between 0 and 1, and then transformed into

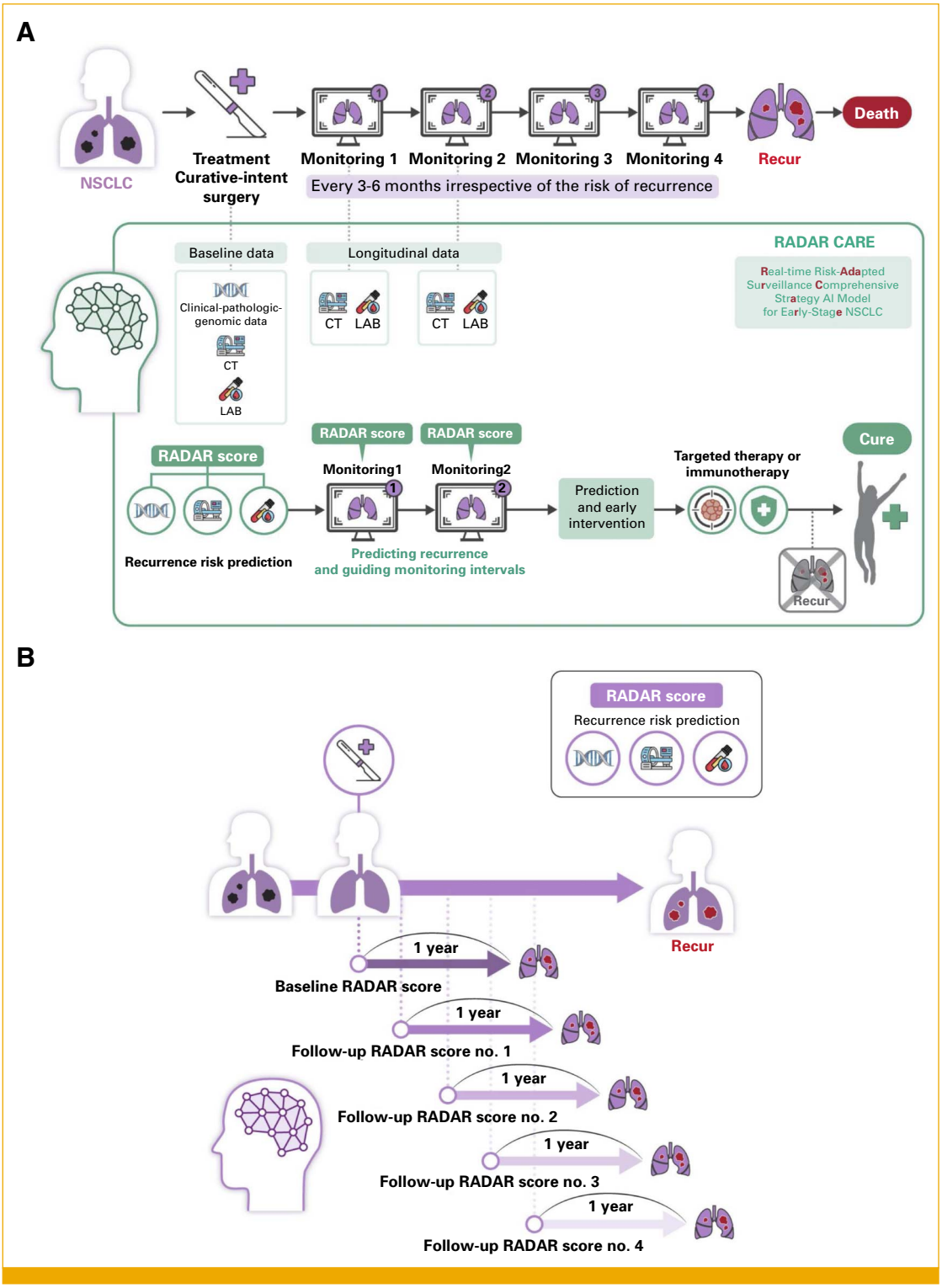

FIG 1. Study scheme of the RADAR CARE study. (A) Study design concept. (B) Definition of the RADAR score. CT, computed tomography; NSCLC, non–small cell lung cancer; RADAR CARE, Real-time Risk-Adapted Surveillance Comprehensive Strategy AI Model for Early-Stage NSCLC.

52-dimensional vectors. The CT imaging results collected were text interpretations from the radiology specialist and thus natural language data. These results were embedded into 768-dimensional vectors using a pretrained natural language processing model, ClinicalBERT.[13,14] Unlike general language models, ClinicalBERT is specifically designed to process clinical data, including medical terminology. ClinicalBERT comprises 12 transformer[15] encoders, each of which

incorporates 12 attention heads. The transformed data were organized into time series data spanning 2 years, with a 30-day interval between timestamps. Each data set was composed of 24 timestamps, with blood test results in a (24, 52) format and CT imaging interpretation results in a (24, 768) format. Clinical and pathological data, as one-off data, were formatted without timestamps as (130). To handle missing values in the longitudinal data (eg, laboratory results and CT imaging interpretation data), we applied the Last Observation Carried Forward method. In this approach, the most recently available value is carried forward to fill subsequent missing time points. This strategy, commonly referred to as forward-filling in time series data processing, helps maintain temporal continuity and reflects typical clinical decision-making patterns, where previous values are often used to estimate a patient's interim status.

Labels were designed to precisely predict the frequency of patient relapse. *Label 1 (Relapse)* was assigned to cases with a relapse within 1 year from the previous timestamp, and *Label 0 (Non-Relapse)* was assigned to cases that did not relapse within 1 year. This method focused on accurately classifying the frequency of relapse to enhance the utility of the model in real clinical settings.

Development of the Multimodal Transformer Model

The input to the model was configured in a multimodal form, consisting of time series data in the shape of (batch_size, timestamp, dimension) and single-point data in the shape of (batch_size, dimension; Data Supplement, Fig S1). The time series data are converted into a single vector in the form (batch_size, dimension) after they pass through a transformer with two encoders. The single-point data are converted into a single vector in the form (batch_size, dimension) after they pass through a fully connected layer. Those singular vectors are then combined into one vector, which passes through a fully connected layer to produce the final output that predicts the likelihood of relapse within 1 year.

Considering the unique characteristics of each input data set, the complexity of the model was adjusted to facilitate effective learning. To process the relatively simple single-point data, a fully connected layer was used to reduce model complexity. In contrast, to process the time-stamped data of size (24, 768), a transformer model with two encoders was used. Each encoder contains 12 attention heads and a feed-forward network of size 1,536, allowing the model to handle complex data patterns. For smaller data of size (24, 52), a transformer model with two encoders, each with four attention heads, and a feed-forward network of size 128 were used to maintain the model's relatively simple structure.

By optimizing the model configuration to reflect the unique modality of each data type, it was possible to integrate the characteristics of various input data effectively and enhance the accuracy of relapse prediction.

Training and Validation

An independent optimization process was conducted for each type of input data by temporarily fixing the model parameters of two of the three inputs while the model learned from the other data type. This process ensured that the unique characteristics of each type of input data were adequately reflected and maximized the overall model learning efficiency and performance optimization.

Different numbers of epochs were assigned for training of each type of input data. The CT imaging interpretation data, laboratory data, and clinical and pathological data were trained for 463, 73, and 21 epochs, respectively. That differentiated learning approach played a crucial role in understanding the unique characteristics of each data type and deriving the optimal model performance.

For the training process, the binary cross-entropy loss function was used. To optimize learning by addressing data imbalance, a higher weight was assigned to the loss function of relapse cases, which are less frequent than nonrelapse cases.

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[\omega_{pos}\cdot y_i \log(\hat{y}_i) + \omega_{neg}\cdot(1-y_i)\log(1-\hat{y}_i)\right]$$

$$(\omega_{pos} = 15.087, \omega_{neg} = 0.517)$$

where N represents the number of samples, $y_i \in \{0, 1\}$ represents the actual label of the ith sample (relapse within 1 year), $\hat{y}_i$ represents the model's predicted probability for the ith sample, ω_{pos} represents the weight for the positive class (Label 1), and ω_{neg} represents the weight for the negative class (Label 0). Adjusting the weights of the loss function enhanced the model sensitivity to relapse cases.

The learning rate was set at 1e-4, with AdamW used as the optimizer and a batch size of 500. The entire code was written using Python v3.7, and the machine-learning algorithm was implemented using TensorFlow v2.5.0.

Statistical Analysis

The RADAR score, calculated on the basis of data at the time of surgery (baseline RADAR score) and during the follow-up period (longitudinal monitoring RADAR score), predicted recurrence within 1 year (Fig 1B). The receiver operating characteristics (ROC) curve and area under the ROC curve (AUC) were used to evaluate the predictive ability of the RADAR score. The optimal threshold was determined using the Youden index and the point closest-to-(0,1). The performance measures for the selected optimal threshold were sensitivity, specificity, precision, and F1 score. To consider the correlations of RADAR scores within patients, the generalized estimating equation method was applied to model the longitudinal monitoring data. The DFS was defined as recurrence within 1 year or death. The optimal cutoff value of

TABLE 1. Baseline Characteristics

Variables	Training Set (n = 11,341)	Test Set (n = 2,836)	*P*
Age, median (range)	64 (13-91)	64 (25-88)	.343
Sex, No. (%)			.436
Male	6,480 (57.1)	1,597 (56.3)	
Female	4,861 (42.9)	1,239 (43.7)	
Pathologic stage, No. (%)			.177
I	8,014 (70.7)	2,054 (72.4)	
II	1,935 (17.0)	451 (15.9)	
III	1,392 (12.3)	331 (11.7)	
ECOG PS, No. (%)			.695
Score 0	6,646 (58.6)	1,646 (58.0)	
Score 1	2,695 (23.8)	644 (22.7)	
Score 2	176 (1.6)	52 (1.8)	
Score 3	18 (0.1)	5 (0.2)	
Smoking, No. (%)			.755
Never-smoker	5,514 (48.6)	1,382 (48.7)	
Ex-smoker	3,105 (32.0)	759 (26.8)	
Current smoker	2,722 (27.4)	695 (24.5)	
Location, No. (%)			.053
Upper lobe, lung	6,098 (53.8)	1,541 (54.3)	
Lower lobe, lung	4,219 (37.2)	1,090 (38.4)	
Middle lobe, lung	750 (6.6)	152 (5.4)	
Lung NOS	193 (1.7)	37 (1.3)	
Overlapping lung lesion	58 (0.5)	14 (0.5)	
Main bronchus	22 (0.2)	2 (0.1)	
EGFR, No. (%)			.779
Negative	5,096(44.9)	1,286(45.3)	
Positive	4,132(36.4)	1,028(36.2)	
ALK, No. (%)			.865
Negative	8,932 (78.8)	2,257 (79.6)	
Positive	258 (2.3)	63 (2.2)	
PD-L1 (TPS, 22C3), No. (%)			.671
Nonexpressor (<1%)	3,493 (30.8)	830 (29.3)	
Expressor (1%-49%)	1,116 (9.8)	283 (10.0)	
High expressor (>49%)	869 (7.7)	215 (7.6)	
PD-L1 (TPS, SP263), No. (%)			.669
Nonexpressor (<1%)	582 (5.1)	136 (4.8)	
Expressor (1%-49%)	301 (2.7)	76 (2.7)	
High expressor (>49%)	140 (1.2)	39 (1.4)	
Histology, No. (%)			.106
AD	8,595 (75.8)	2,136 (75.3)	
SQ	2,099 (18.5)	528 (18.6)	
Other	647 (5.7)	172 (6.1)	
Differentiation, No. (%)			.528
WD	1,043 (9.2)	281 (9.9)	
MD	7,392 (65.2)	1,848 (65.2)	
PD	1,772 (15.6)	442 (15.6)	
Unclassified	1,134 (10.0)	265 (9.3)	
Pleural invasion, No. (%)			.390
PL0	9,185 (81.0)	2,305 (81.3)	

(continued on following page)

TABLE 1. Baseline Characteristics (continued)

Variables	Training Set (n = 11,341)	Test Set (n = 2,836)	*P*
PL1	650 (5.7)	185 (6.5)	
PL2	524 (4.6)	115 (4.1)	
PL3	289 (2.5)	69 (2.4)	
Not identified	16 (0.1)	4 (0.1)	
Pattern, No. (%)			.718
Lepidic	1,857 (16.4)	465 (16.4)	
Acinar	6,401 (56.4)	1,617 (57.0)	
Papillary	2,049 (18.1)	495 (17.5)	
Micropapillary	804 (7.1)	181 (6.4)	
Solid	1,207 (10.6)	297 (10.5)	
Mucinous	856 (7.5)	200 (7.1)	
Colloid	12 (0.1)	1 (0.04)	
Fetal	10 (0.09)	2 (0.07)	
Basaloid	35 (0.3)	13 (0.5)	
Neuroendocrine	194 (1.7)	51 (1.8)	
Large cell	180 (1.6)	46 (1.6)	
Pleomorphic	157 (1.4)	42 (1.5)	
Spindle cell	50 (0.4)	20 (0.7)	
Giant cell	44 (0.4)	13 (0.5)	
Clear cell	21 (0.2)	6 (0.2)	
NUT	3 (0.03)	1 (0.04)	
Mucoepidermoid	35 (0.3)	7 (0.2)	
Adenoid cyst	9 (0.08)	4 (0.1)	
Cribriform	248 (2.2)	60 (2.1)	
Epithelial myoepithelial	7 (0.06)	1 (0.04)	

Abbreviations: AD, adenocarcinoma; ALK, anaplastic lymphoma kinase; ECOG PS, Eastern Cooperative Oncology Group performance status; EGFR, epidermal growth factor receptor; MD, moderate differentiation; NOS, not otherwise specified; NUT, NUT carcinoma; PD, poorly differentiation; SQ, squamous cell carcinoma; TPS, tumor proportion score; WD, well deferentiation.

the RADAR score for DFS was determined using the Contal and O'Quigley method,[16] and the difference between risk groups is presented with Kaplan-Meier curves and a Cox proportional hazard model. All statistical analyses were performed using SAS software (version 9.4) and R software (version 4.2.2; R Foundation for Statistical Computing).

Ethics Statement

This study was reviewed and approved by the Institutional Review Board of Samsung Medical Center (No. 2023-12-001).

RESULTS

Study Population and Clinical Characteristics

Of the 14,618 patients who were diagnosed with NSCLC and received surgical resection with curative intent at Samsung Medical Center between January 2008 and September 30, 2022 (Data Supplement, Fig S2), 444 were excluded: 401 had double primary cancer, three had primary lung cancer, 35 underwent R1 resection, and two were excluded due to missing longitudinal data. The final study population comprised 14,177 patients with available baseline clinical-pathologic-genomic data and longitudinal data. Data cutoff was February 2024, and the median follow-up duration was 59.0 months (range, 0.5-190.1). Table 1 shows the baseline characteristics. The median age at the time of surgery was 64 years (range, 13-91), 57.2% of patients were male, and 48.8% of patients were never-smokers. The pathologic stage was I, II, and IIIA in 71.3%, 16.9%, and 12.2% of patients, respectively. The model incorporates 64 clinical-pathological-molecular factors at baseline along with longitudinal laboratory and CT imaging interpretation data (Data Supplement, Table S1). In total, 177,246 CT imaging interpretations were used (mean 12.4 chest CT scans per patient) during surveillance.

The data collected from the 14,177 patients were randomly split into training and validation sets in an 8:2 ratio (11,341 in the training set and 2,836 in the validation set). The training data contained 2,306 relapse cases (20.3%) and 9,035 nonrelapse cases (79.7%), and the validation data contained 568 relapse cases (20%) and 2,268 nonrelapse cases (80%).

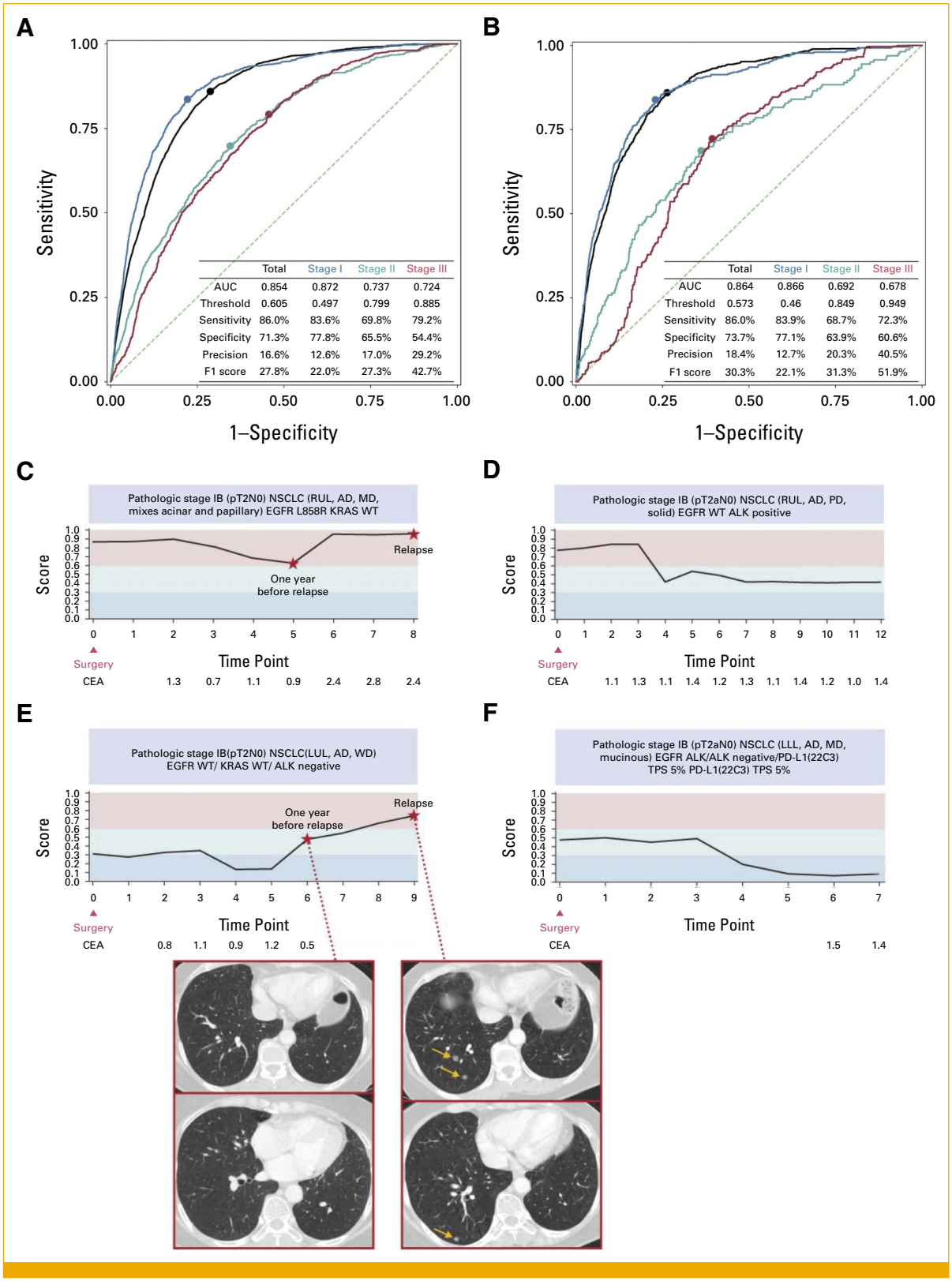

FIG 2. (A) ROC curve of the baseline and longitudinal RADAR score for recurrence within 1 year in the total population and in subgroups according to pathologic stage. (B) ROC curve of the baseline and longitudinal RADAR score for recurrence within 1 year in patients with *EGFR* mutation–positive NSCLC and in subgroups according to pathologic stage. (C) A case with a high baseline RADAR score (>0.6) and persistently elevated scores during longitudinal monitoring. (D) A case with a high initial RADAR score (>0.6) that declined during longitudinal follow-up. (E) A case with an initially low RADAR score (≤0.3) that showed an upward trend during longitudinal monitoring. (F) A case with a persistently low RADAR score (≤0.3). ALK, anaplastic lymphoma kinase; AD, adenocarcinoma; CEA, carcinoembryonic antigen; EGFR, epidermal growth factor receptor; (continued on following page)

FIG 2. (Continued). LUL, left upper lobe; MD, moderate differentiation; NSCLC, non–small cell lung cancer; PD, poorly differentiation; ROC, receiver operating characteristics; RUL, right upper lobe; TPS, tumor proportion score; WD, well differentiation.

Baseline RADAR Score of Prediction Model

The mean baseline RADAR score in the total population was 0.436 (standard deviation [SD], 0.302). According to the pathologic stage, the mean baseline RADAR score was 0.183 (SD, 0.190), 0.564 (SD, 0.217), and 0.712 (SD, 0.192) for stages I, II, and III, respectively (P <.001). There was statistically significant difference of baseline RADAR score between *EGFR* mutation–positive NSCLC and *EGFR*-wild NSCLC (mean baseline RADAR score was 0.387 [SD, 0.301] for *EGFR* mutation–positive NSCLC *v* 0.529 [SD, 0.283] for *EGFR* wild-type NSCLC, P <.0001). The baseline RADAR score predicted 1-year recurrence with an AUC of 0.823. According to the pathologic stage, the baseline RADAR score predicted recurrence with AUCs of 0.844, 0.618, and 0.683 for stages I, II, and IIIA, respectively.

Predictive Model With Baseline and Longitudinal RADAR Score

Figure 2 shows the AUC of the baseline and longitudinal RADAR score for recurrence within 1 year in the total population and in subgroups according to pathologic stage. In the total population, the baseline and longitudinal RADAR score predicted 1-year recurrence with an AUC of 0.854 (Fig 2A). According to pathologic stage, the baseline and longitudinal RADAR scores predicted recurrence with AUCs of 0.872, 0.737, and 0.724 for stages I, II, and IIIA, respectively. In patients with *EGFR* mutation, the baseline and longitudinal RADAR score predicted 1-year recurrence with an AUC of 0.864 (Fig 2B). According to the pathologic stage for patients with *EGFR* mutation, the baseline and longitudinal RADAR score predicted recurrence with AUCs of 0.866, 0.692, and 0.678 for stages I, II, and IIIA, respectively. The Data Supplement (Table S2) shows the relapse rate according to the cutoff value for the RADAR score. In the total population, the odds ratio of relapse occurring within 1 year of the relevant follow-up time increases by 1.27 times for each 0.1 increase in the RADAR score (95% CI, 1.24 to 1.29, P <.0001). In subgroups with disease in pathologic stages I, II, and III, a RADAR score increase of 0.1 increased the occurrence of relapse by an odds ratio [95% CI] of 1.32 [1.28 to 1.37], 1.08 [1.06 to 1.11], and 1.06 [1.03 to 1.09], respectively (P <.001 for all).

Four Trends in the Baseline and Longitudinal Monitoring RADAR Score

The baseline and longitudinal monitoring RADAR score predicted disease recurrence with an AUC of 0.854. Figure 2A shows the AUC, threshold, sensitivity, specificity, precision, and F1 score. The patterns in the baseline and longitudinal

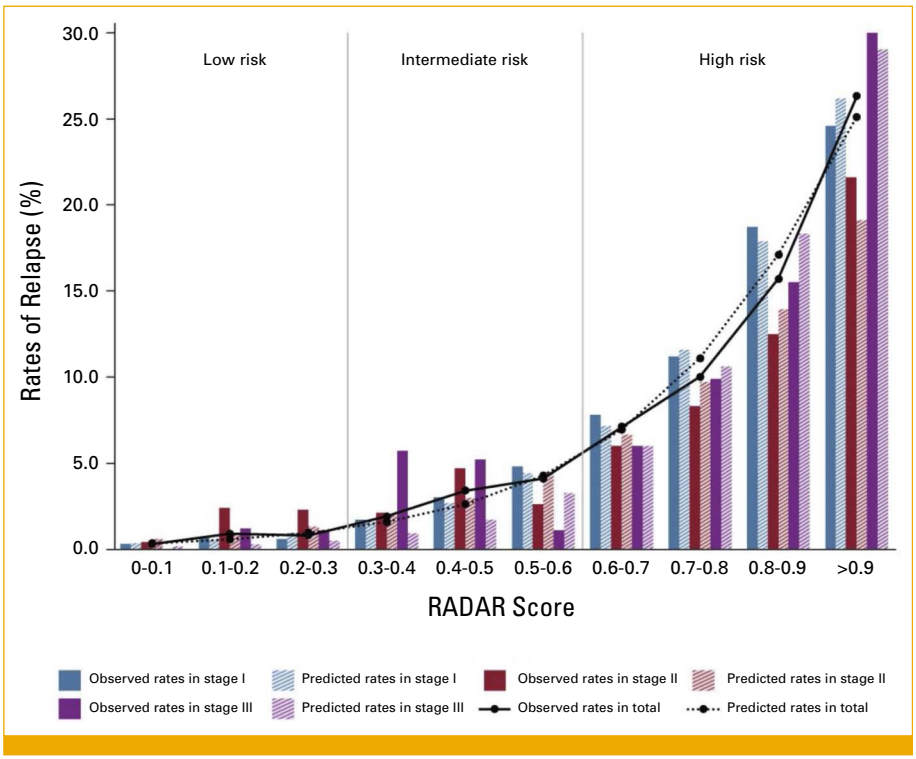

FIG 3. The relationship between RADAR score and relapse within 1 year.

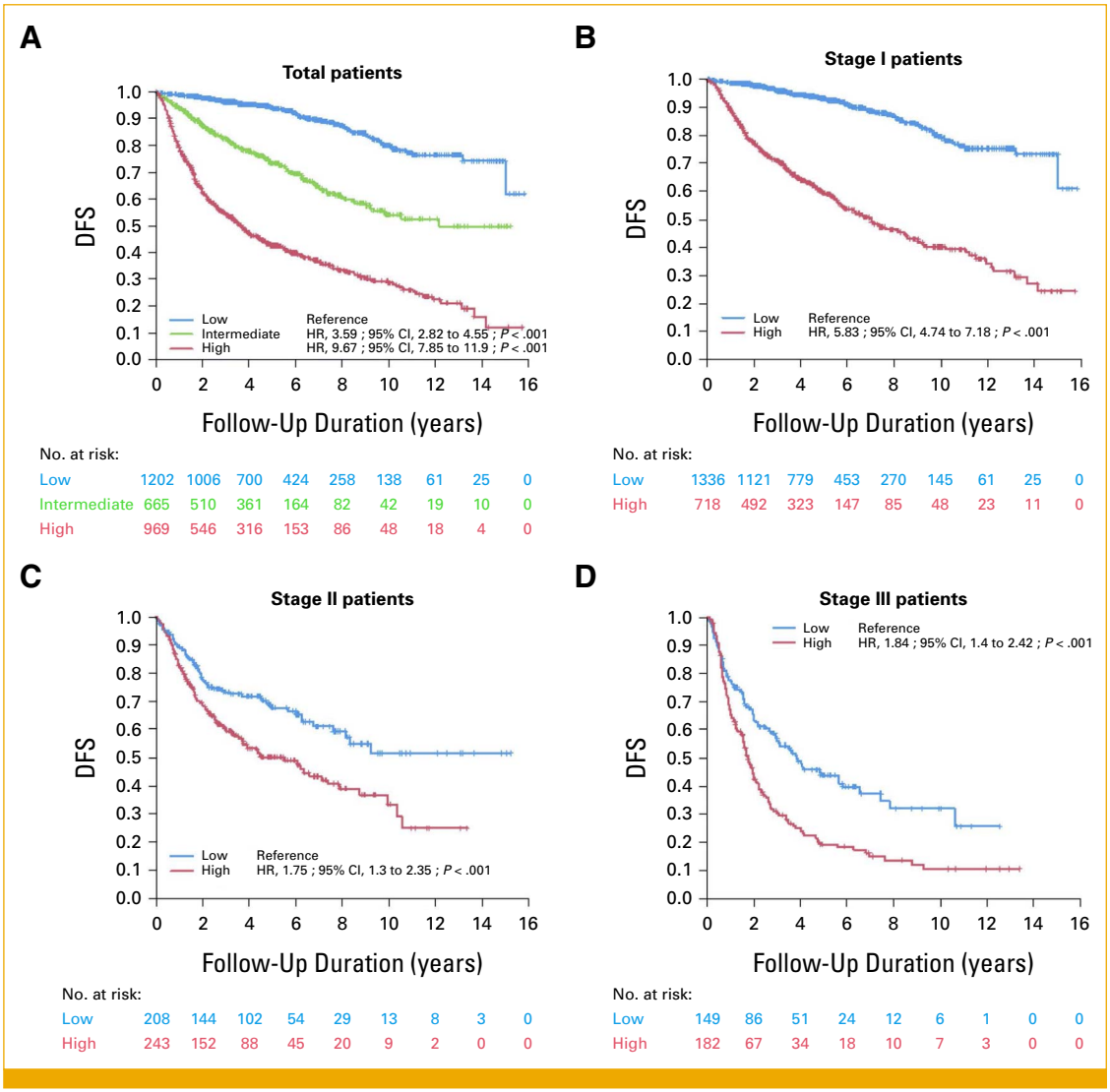

FIG 4. (A) Survival curves for DFS according to baseline RADAR score and pathologic stage in the total population. (B) Survival curves for DFS in stage I patients. (C) Survival curves for DFS in stage II patients. (D) Survival curves for DFS in stage III patients. DFS, disease-free survival; HR, hazard ratio.

monitoring RADAR scores were classified into four types: (1) high baseline RADAR score (>0.6) and persistently high RADAR score during longitudinal monitoring, (2) high baseline RADAR score (>0.6) and decreased RADAR score during longitudinal monitoring, (3) low (≤0.3) baseline RADAR score and increased RADAR score during longitudinal monitoring, and (4) persistently low (≤0.3) RADAR score. Figures 2C-2F shows a typical case of each pattern. The Data Supplement (Table S3) shows risk-stratified surveillance and adjuvant treatment strategies for each RADAR score pattern.

Threshold for Baseline RADAR and DFS

The calibration plot presented in Figure 3 and the Data Supplement (Table S4) assesses the agreement between observations and RADAR score predictions of relapse within 1 year. The predictions were confirmed to be accurate based on comparison with the observed values. On the basis of RADAR scores of 0.3 and 0.6, patients were categorized into low-, intermediate-, and high-risk groups. These thresholds were selected based on the observed 1-year relapse rates and their clinical significance: Patients with a RADAR score <0.3 had a 1-year relapse rate of <1%. Patients with scores between 0.3 and 0.6 had a relapse rate of <5%. Patients with scores ≥0.6 had a relapse rate of 5% or higher. As the RADAR score increased, the relapse rate within 1 year increased, and that trend was consistent across pathological stages. In particular, even if it was an early stage, a high RADAR score was observed to have a high relapse rate within 1 year, and even if it was an advanced stage, a low RADAR score was observed to have a low relapse rate within 1 year. This demonstrated that the RADAR score predicts relapse within 1 year regardless of the pathological stage.

Figure 4 presents DFS according to risk group in the total population and patients in pathologic stages I, II, and III. In the total population, the risk groups were categorized using the first RADAR score for each patient. A relapse or death within 1 year was significantly more likely in the intermediate- and high-risk groups, with hazard ratios (HRs) of 3.59 (95% CI, 2.82 to 4.55) and 9.67 (95% CI, 7.85 to 11.90), respectively, compared with low-risk patients (P <.001). In pathologic stages I, II, and III, the occurrence of relapse or death within 1 year increased significantly in the high-risk group compared with the low-risk group (HR [95% CI], 5.83 [4.74 to 7.18], 1.75 [1.30 to 2.35], and 1.84 [1.40 to 2.42] in stages I, II, and III, respectively; P < .001 for all).

DISCUSSION

In this pilot study, a model for predicting recurrence during surveillance was successfully developed. The RADAR score, a risk score for recurrence at the time of surgery and during surveillance, incorporates radiologic and laboratory data. The prediction power of the baseline data showed an AUC of 0.823, and the prediction power of the longitudinal data presented an AUC of 0.766, which increased to 0.854 when the baseline and longitudinal data were used together. This study further suggests risk-adapted treatment based on the baseline and longitudinal monitoring RADAR score.

The multimodal model developed in this study was designed to efficiently integrate and optimize different types of input data[17,18]: clinical and pathological data, laboratory data, and CT imaging interpretation data. Furthermore, a transformer model was used to precisely analyze the interactions among the data points over time. The multimodal learning method successfully reflects the unique information inherent in each data type, and the transformer model enhanced the precision with which the interactions among data points over time were analyzed.

The major strength of this study is design of a model based on clinical, pathologic, and molecular baseline data at the time of surgery and longitudinal monitoring data collected in routine clinical practice using a significantly larger cohort of 14,177 patients. By using a transformer-based architecture, RADAR effectively captured dynamic patterns over time, resulting in a higher AUC of 0.854. Previous predictive model relied on static, single time point data and did not incorporate temporal or longitudinal information.[19] These distinctions highlight RADAR's potential as a more robust and clinically applicable tool for real-time recurrence risk prediction. The predictive power is impressive, especially for pathologic stage I, which had a low recurrence rate, indicating that the RADAR score can effectively predict recurrence in a subgroup with few recurrence events. The larger number of stage I cases compared with stages II and IIIA also contributed to this finding. This pilot study showed four trends in the baseline and longitudinal RADAR scores. For patients with a high baseline RADAR score (>0.6) and persistently high RADAR score during longitudinal monitoring, aggressive adjuvant treatment such as immunotherapy or targeted therapy is needed immediately after surgery (Data Supplement, Table S3). Patients with the second trend, a high baseline RADAR score (>0.6) and decreased RADAR score during longitudinal monitoring, need aggressive adjuvant treatment after surgery, but the duration of treatment can be adjusted according to the trend in the RADAR score. Patients with the third trend, a low (≤0.3) baseline RADAR score that increases during longitudinal monitoring, do not need aggressive adjuvant treatment after surgery, but additional treatment should be applied to prevent recurrence when the RADAR score increases. Patients with the fourth trend of persistently low (≤0.3) RADAR scores require less frequent follow-up monitoring than for other groups.

Despite its strengths, this study also has limitations. The model includes only CT imaging interpretation data from radiologists and not the original image data. In addition, the RADAR score needs to be validated using multicenter and multinational data. A further prospective study to test the suggested risk-adapted treatment strategy using the RADAR score is warranted.

In conclusion, in this proof-of-concept study, the RADAR score successfully predicted disease recurrence 1 year in advance in patients with early-stage NSCLC who received surgical resection with curative intent. Training and internally validating the RADAR score with an independent data set improved its prediction power. To our knowledge, this is the first study to propose a multimodal transformer machine-learning model to predict recurrence in oncology. Patients with high RADAR score had a higher incidence of recurrence in 1 year and shorter DFS than those who had low RADAR score. This suggests that RADAR score could serve as a complementary biomarker for recurrence. A further prospective study of the clinical implementation of the RADAR score in treating early-stage NSCLC is warranted.

AFFILIATIONS

[1]Division of Hematology-Oncology, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[2]Spidercore Inc, Daejeon, Republic of Korea
[3]Biomedical Statistics Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea
[4]Department of Medical Education and Medical Humanities, College of Medicine, Inha University, Incheon, Republic of Korea
[5]Department of Radiology and Center for Imaging Science, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[6]Department of Thoracic and Cardiovascular Surgery, Sungkyunkwan University School of Medicine, Samsung Medical Center, Seoul, Republic of Korea

CORRESPONDING AUTHOR

Hong Kwan Kim, MD, PhD; Social Media Handle: @HyunAeJung62662; e-mail: hkts@skku.edu.

EQUAL CONTRIBUTION

H.A.J., D.L. and B.P. contributed equally to this work.

SUPPORT

Supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT; IRIS RS-2025-00521527).

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/po/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Hyun Ae Jung
Consulting or Advisory Role: Yuhan, Guardant Health, AIMEDBIO
Research Funding: Yuhan

Ho Yun Lee
Stock and Other Ownership Interests: Cirelinesoft Co
Consulting or Advisory Role: Roche
Research Funding: Johnson & Johnson Inc, Lunit, Bracco Diagnostics, AstraZeneca

Jong-Mu Sun
Research Funding: Yuhan (Inst)

Se-Hoon Lee
Honoraria: AstraZeneca/MedImmune, Roche, Lilly, Amgen, Yuhan, Merck Sharp & Dohme, Bristol Myers Squibb Foundation
Consulting or Advisory Role: AstraZeneca, Roche, Pfizer, Lilly, BMS/Ono, Takeda, Janssen, IMBdx, Abion, BeiGene, Daiichi Sankyo, ImmuneOncia, Merck (German), Merck Sharp & Dohme, Novartis
Speakers' Bureau: Abion
Research Funding: AstraZeneca (Inst), Lunit (Inst), Merck Sharp & Dohme (Inst), Daiichi Sankyo

Jin Seok Ahn
Honoraria: Roche, BC World Pharmaceutical, Yuhan, Novartis, Amgen, Boehringer Ingelheim, Menarini, AstraZeneca, Takeda, Boryung, Samyang, Nokwon Medical, BMS Korea, LG Chem, Daiichi Sankyo Korea
Consulting or Advisory Role: Pharmbio Korea, Yuhan, ImmuneOncia, Daiichi Sankyo Korea, Roche, Pfizer

Myung-Ju Ahn
Honoraria: AstraZeneca, Lilly, MSD, Takeda, Amgen, Merck Serono, Yuhan, Daiichi Sankyo/Astra Zeneca
Consulting or Advisory Role: AstraZeneca, Lilly, MSD, Takeda, Alpha Pharmaceutical, Amgen, Merck Serono, Pfizer, Yuhan, Arcus Ventures, Daiichi Sankyo/Astra Zeneca, Daiichi Sankyo/Astra Zeneca
Research Funding: Yuhan

No other potential conflicts of interest were reported.

AUTHOR CONTRIBUTIONS

Conception and design: Hyun Ae Jung, Daehwan Lee, Ho Yun Lee, Tae Jung Kim, Yong Soo Choi, Se-Hoon Lee, Myung-Ju Ahn, Hong Kwan Kim
Administrative support: Hyun Ae Jung, Hong Kwan Kim
Provision of study materials or patients: Tae Jung Kim, Jong-Mu Sun, Jin Seok Ahn
Collection and assembly of data: Hyun Ae Jung, Daehwan Lee, Ho Yun Lee, Tae Jung Kim, Yeong Jeong Jeon, Junghee Lee, Seong Yong Park, Jong Ho Cho, Yong Soo Choi, Sehhoon Park, Jong-Mu Sun, Se-Hoon Lee, Jin Seok Ahn, Myung-Ju Ahn, Hong Kwan Kim
Data analysis and interpretation: Hyun Ae Jung, Daehwan Lee, Boram Park, Kiwon Lee, Ho Yun Lee, Tae Jung Kim, Jong Ho Cho, Yong Soo Choi, Sehhoon Park, Se-Hoon Lee, Jin Seok Ahn, Myung-Ju Ahn, Hong Kwan Kim
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

ACKNOWLEDGMENT

The authors express their gratitude to Seung-Ho Shin, a valued member of our ROOT-HEALTH team since 2018, for his contributions to this work through data extraction as a back-end developer. Additionally, we extend our thanks to Hyeon Yoo, who has served as the data manager of the ROOT-HEALTH team since 2020.

REFERENCES

1. National Cancer Center, South Korea: Domestic Cancer Registry Project Report. 2020. https://www.cancer.go.kr/lay1/S1T648C650/contents.do
2. Planchard D: Adjuvant osimertinib in EGFR-mutated non-small-cell lung cancer. N Engl J Med 383:1780-1782, 2020
3. 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 11:39-51, 2016
4. Cao W, Tang Q, Zeng J, et al: A review of biomarkers and their clinical impact in resected early-stage non-small-cell lung cancer. Cancers (Basel) 15:4561, 2023
5. Wu YL, Tsuboi M, He J, et al: Osimertinib in resected EGFR-mutated non-small-cell lung cancer. N Engl J Med 383:1711-1723, 2020
6. Wu YL, Dziadziuszko R, Ahn JS, et al: Alectinib in resected ALK-positive non-small-cell lung cancer. N Engl J Med 390:1265-1276, 2024
7. Felip E, Altorki N, Zhou C, et al: Adjuvant atezolizumab after adjuvant chemotherapy in resected stage IB-IIIA non-small-cell lung cancer (IMpower010): A randomised, multicentre, open-label, phase 3 trial. Lancet 398:1344-1357, 2021
8. Park S, Ock CY, Kim H, et al: Artificial intelligence-powered spatial analysis of tumor-infiltrating lymphocytes as complementary biomarker for immune checkpoint inhibition in non-small-cell lung cancer. J Clin Oncol 40:1916-1928, 2022
9. Mikhael PG, Wohlwend J, Yala A, et al: Sybil: A validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. J Clin Oncol 41:2191-2200, 2023
10. Chi SA, Yu H, Choi YL, et al: Trends in survival rates of non-small cell lung cancer with use of molecular testing and targeted therapy in Korea, 2010-2020. JAMA Netw Open 6:e232002, 2023
11. Jung HA, Lee DH, Lim SM, et al: Epidemiology and outcomes of non-small cell lung cancer in South Korea. JAMA Netw Open 7:e2355331, 2024
12. Jung HA, Noh JM, Sun JM, et al: Real world data of durvalumab consolidation after chemoradiotherapy in stage III non-small-cell lung cancer. Lung Cancer 146:23-29, 2020
13. Devlin J, Chang MW, Lee K, et al: Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. https://arxiv.org/abs/1810.04805
14. Huang K, Altosaar J, Ranganath R: Clinicalbert: Modeling clinical notes and predicting hospital readmission. 2019. https://arxiv.org/abs/1904.05342
15. Vaswani A, Shazeer N, Parmar N, et al: Attention is all you need. 2017. https://arxiv.org/abs/1706.03762

16. Contal C, O'Quigley J: An application of changepoint methods in studying the effect of age on survival in breast cancer. Comput Stat Data Anal 30:253-270, 1999
17. Harutyunyan H, Khachatrian H, Kale DC, et al: Multitask learning and benchmarking with clinical time series data. Sci Data 6:96, 2019
18. Zhou HY, Yu Y, Wang C, et al: A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nat Biomed Eng 7:743-755, 2023
19. Janik A, Torrente M, Costabello L, et al: Machine learning-assisted recurrence prediction for patients with early-stage non-small-cell lung cancer. JCO Clin Cancer Inform 10.1200/CCI.22.00062

