CANCER

Radiotranscriptomics in papillary thyroid carcinoma complement current noninvasive risk stratification system

Dong Hyun Seo[1]†, Eunjung Lee[2]†, Jung Hyun Yoon[3]†, Eun Gyeong Park[1]†, Sunmi Park[1], Hwa Young Lee[4], Joon Ho[5], Cho Rok Lee[5], Kyunghwa Han[3], Jandee Lee[4]*, Jin Young Kwak[3]*, Young Suk Jo[1]*

Papillary thyroid carcinoma (PTC) generally has a favorable prognosis; however, overtreatment persists because of the lack of reliable noninvasive risk stratification tools. This study developed a radiomics-based approach to enhance the preoperative assessment of PTC. Imaging features from 255 patients were analyzed, and three tumor clusters were identified via unsupervised clustering, with one cluster (Cluster 2) displaying favorable clinical and molecular profiles. A radiomics score was constructed and validated internally and externally, achieving high diagnostic accuracy (area under the curve of 0.98) and independently predicting benign features such as a lower N stage and favorable treatment responses. Transcriptomic analysis revealed immune activation and survival-related gene expression in Cluster 2. The model demonstrated robust performance in stratifying patients for active surveillance and may complement current diagnostic frameworks, offering a precise, noninvasive tool to guide clinical decision-making.

INTRODUCTION

Although the global incidence of thyroid cancer has been on the rise, in some countries, it has stabilized, largely due to widespread high-quality imaging refining detection practices (*1*). Advancements in ultrasound (US) techniques have considerably increased the early detection of cancer, particularly in the microcarcinoma stage (*2*). However, 90% of patients with thyroid cancer are diagnosed with papillary thyroid carcinoma (PTC), which is associated with a favorable prognosis and a 5-year survival rate of 90 to 99% (*3*–*5*). This high rate of early detection can sometimes lead to overtreatment, as many cases of PTC may not require aggressive intervention given their favorable survival outcomes.

Considering these clinical outcomes, overtreatment remains a notable concern for clinicians, leading to the widespread acceptance of active surveillance (AS) for low-risk PTC, particularly papillary thyroid microcarcinoma (<1 cm) (*6*, *7*). Recent studies, including those by Altshuler *et al.* and Ho *et al.*, have broadened the scope of AS for low-risk PTC, demonstrating its safety for nodules ≥2 cm, with minimal progression and effective rescue surgery when necessary (*8*, *9*). However, accurately predicting which nodules will progress remains a challenge, highlighting the importance of advanced tools, such as radiomics, for improved risk stratification in AS candidates (*2*, *10*).

Extensive bioinformatics analyses have defined different molecular subtypes of PTC and uncovered numerous genomic markers associated with aggressive prognosis (*11*–*13*). Among these, the *BRAF*^{V600E} mutation and telomerase reverse transcriptase (*TERT*) promoter mutations are two renowned genomic markers that are closely linked to increased mortality (*14*–*16*). Developing feasible and noninvasive methods to predict such aggressive genomic indicators is of important value in advancing precision medicine.

Lymph node metastasis (LNM), occurring in 30 to 70% of patients with PTC, is a crucial prognostic factor (*5*, *17*) owing to its strong association with poor outcomes, including reduced survival and increased metastasis, which markedly influence treatment decisions (*18*, *19*). However, some studies have suggested that the prognostic value of LNM staging, particularly in central neck node metastasis, remains unclear in terms of survival or disease progression (*20*–*22*). Regarding tumor heterogeneity, classifications of LNM based on distinct tumorigenic characteristics and varying outcomes have been proposed, raising doubts about the prognostic impact of LNM (*23*, *24*). This highlights the need for alternative diagnostic tools to accurately identify "high-risk" patients with LNM, as those with suspected LNMs are recommended for invasive examinations (*10*). However, previous artificial intelligence (AI) models for thyroid cancer have primarily focused on predicting LNM and have demonstrated noteworthy performance using radiomics across US, computed tomography, and magnetic resonance imaging (*25*–*27*).

In this study, we introduced three novel radiomics-assisted unsupervised clusters with distinct biological and clinical outcomes. We identified one specific cluster (Cluster 2) with notably favorable characteristics and developed a cluster-specific radiomics scoring system for Cluster 2 using machine learning. Our scoring system significantly identified high-risk patients with PTC, influencing treatment options and predicting recurrence. To interpret the biological relevance of our scoring system, we integrated genomic and transcriptomic analyses, revealing that the Cluster 2 score is significantly associated with RAS-like biology and gene enrichment

[1]Department of Internal Medicine, Open NBI Convergence Technology Research Laboratory, Yonsei University College of Medicine, Seoul 03722, South Korea. [2]School of Mathematics and Computing (Computational Science and Engineering), Yonsei University, Seoul 03722, South Korea. [3]Department of Radiology, Severance Hospital, Research Institute of Radiological Science, Yonsei University College of Medicine, Seoul 03722, South Korea. [4]Department of Surgery, Open NBI Convergence Technology Research Laboratory, Yonsei Cancer Center, Severance Hospital, Yonsei University College of Medicine, Seoul 03722, South Korea. [5]Department of Surgery, Yongin Severance Hospital, Gyeonggi-do 16995, South Korea.
*Corresponding author. Email: jandee@yuhs.ac (J.L.); docjin@yuhs.ac (J.Y.K.); joys@yuhs.ac (Y.S.J.)
†These authors contributed equally to this work.

crucial for cell differentiation and adaptive immune responses. Our findings were independently validated in a dataset of patients with PTC, mostly including those with microcarcinomas, reflecting the actual clinical scenario of low-risk PTC (*28*). Furthermore, the external validation of our data reinforced these results, highlighting the importance of standardizing diagnostic values to facilitate the integration of radiomics into clinical practice.

RESULTS

Study pipeline

A schematic workflow of the research process is shown in Fig. 1. Briefly, our exploratory dataset consisted of retrospectively collected clinicopathological and radiological data from 255 patients with PTC who underwent thyroidectomy between 2014 and 2018 at our institution. For validation purposes, we enrolled independent external data ($n = 203$) and an in-house cohort of patients with small PTC ($n = 150$). The baseline patient characteristics are presented in Table 1.

A total of 730 radiomic features were extracted per US image using in-house texture analysis algorithms. Information on the extracted radiomic features and their intraclass correlation coefficient (ICC) values is provided in data S1.

To investigate the underlying molecular mechanisms of radiomic clusters, we analyzed bulk RNA sequencing data from 255 patients with PTC in our exploratory dataset. For validation, tissue samples and paraffin-embedded slides from the validation cohorts were collected, and immunohistochemistry (IHC) and quantitative polymerase chain reaction (qPCR) were performed on internal and external validation datasets, respectively, to measure gene expression.

Clinicogenomic characterization of radiomics-based unsupervised clusters

To ensure feature robustness, we set an ICC cutoff value of 0.5 and identified 285 reliable radiomic features. We reduced multicollinearity by removing features with a Pearson correlation coefficient of >0.9. Consequently, 75 radiomic features were selected for the downstream analysis (Fig. 2A). Unsupervised clustering using nonmatrix factorization (NMF) was performed, and clustering quality was assessed using the cophenetic coefficient. We identified the optimal number of clusters to be three ($K = 3$), which yielded the highest cophenetic coefficient of 0.9958 (Fig. 2B). A heatmap of

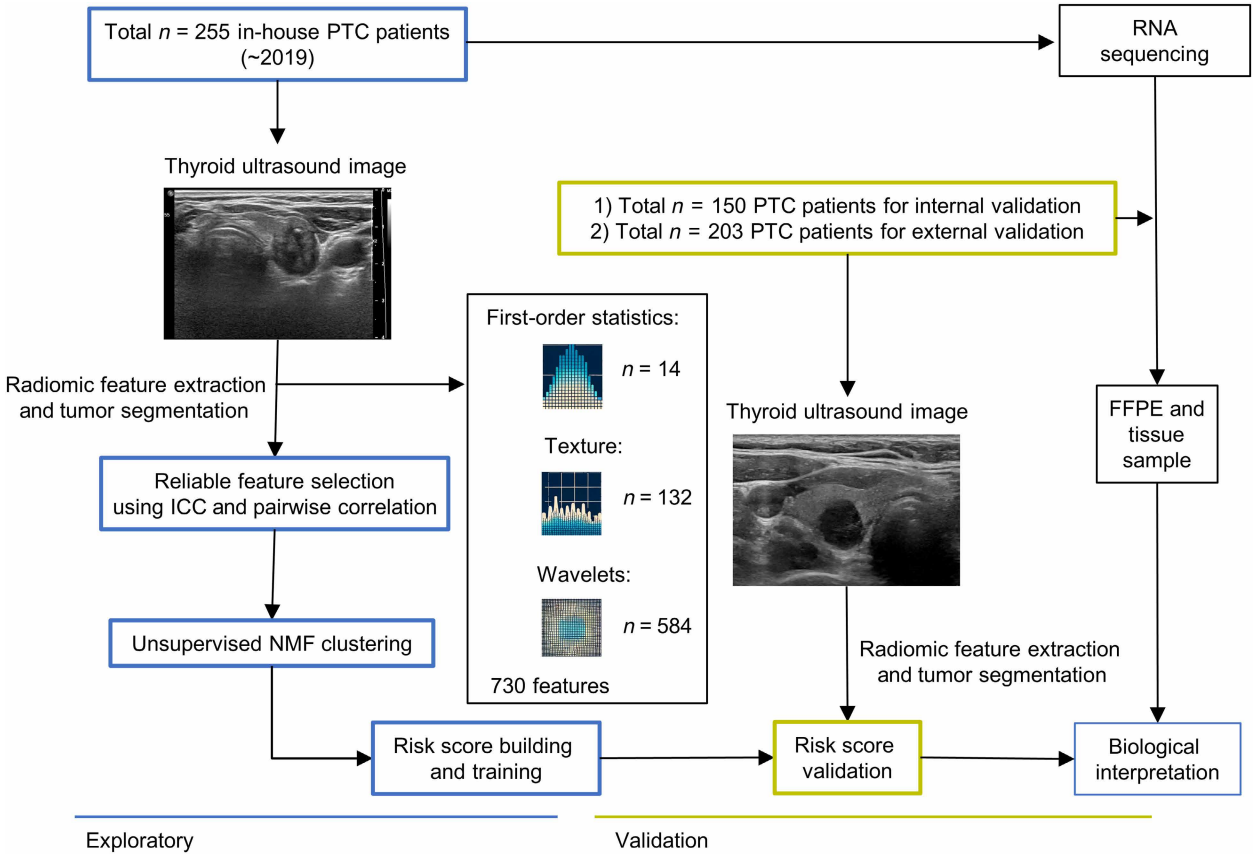


Fig. 1. Workflow pipeline used in the study. In our exploratory study dataset, 255 in-house patients with PTC who underwent surgery before 2019 were enrolled. For validation, 150 independent in-house PTC patients who underwent surgery after 2019 were included. The external validation set consisted of 203 patients with PTC from the Yongin Severance Hospital. In total, 608 patients with PTC were enrolled in this study. Preoperative US images were acquired for each patient, and the cancerous region was segmented by a specialized radiologist. Radiomic features were extracted from the region of interest, followed by a comprehensive feature-selection process to ensure the reproducibility of the findings. A risk score model was built using radiomics and validated in two independent PTC cohorts. RNA sequencing was performed during the exploratory phase to elucidate underlying molecular mechanisms. Subsequently, quantitative polymerase chain reaction (qPCR) and immunohistochemistry (IHC) were used in samples from the validation dataset to validate our molecular mechanisms and assess the clinical significance of our proposed radiomic model. FFPE, formalin-fixed, paraffin-embedded; ICC, intraclass correlation coefficient; NME, nonmatrix factorization.

Table 1. Baseline clinicopathological characteristics of patients with PTC for exploratory and validation sets. Continuous variables are presented as means ± standard deviation. Categorical variables were summarized as the number of patients and their corresponding percentages. TNM stage was classified according to the eighth edition of the American Joint Committee on Cancer Cancer Staging Manual. Significant demographic differences between the patient cohorts were evaluated using one-way analysis of variance (ANOVA) for continuous variables and the chi-square test for proportions of categorical variables. Statistical significance was set at $P < 0.05$. *$P < 0.05$; **$P < 0.01$. NA, not applicable; EBV, Epstein-Barr virus; NTRK, neurotrophic tyrosine receptor kinase; pTERT, *TERT* promoter; RAITx, radioiodine therapy.

Variables	Exploratory data ($n = 255$) (~2018)	Internal validation data ($n = 150$) (2018 ~ 2023)	External validation data ($n = 203$) (2018 ~ 2023)
Age*	50.5 ± 14.4	41.1 ± 12.6	41.4 ± 11.6
Sex			
Male	66 (25.9%)	39 (26%)	42 (20.7%)
Female	189 (74.1%)	111 (74%)	161 (79.3%)
Tumor size (cm)**	1.92 ± 0.85	1.11 ± 0.18	1.30 ± 0.36
T stage**			
T1–2	46 (18%)	137 (91.3%)	75 (36.9%)
T3–4	209 (82%)	13 (8.7%)	128 (63.1%)
N stage**			
N0	82 (32.1%)	48 (32%)	86 (42.3%)
N1a	78 (30.5)	38 (25.3%)	86 (42.3%)
N1b	95 (37.4%)	64 (44.7%)	31 (13.4%)
M stage (M1, distant metastasis)	4 (1.5%)	1 (0.7%)	0
BRAF*^{V600E} mutation	221 (88.4%)	90 (60%)	180 (88.6%)
pTERT mutation*	29 (11.3%)	5 (3.3%)	2 (1.0%)
Gene fusion (EBV, NTRK)	17 (6.6%)	NA	NA
Adjuvant RAITx**			
No RAITx	67 (26.3%)	73 (48.6%)	168 (82.8%)
Low dose (<60 mCi)	49 (19.2%)	10 (6.7%)	1 (0.5%)
Intermediate dose (100 ~ 120 mCi)	46 (18%)	66 (44%)	19 (9.3%)
High dose (>150 mCi)	93 (36.5%)	1 (0.7%)	15 (7.4%)

the normalized values for the 75 radiomic features was plotted to identify three clusters ($K = 3$; Fig. 2C).

To evaluate the clinical relevance of these radiomic clusters, we compared key clinical variables and driver mutation statuses across the clusters, with the results summarized in Table 2. Significant differences were observed in the distribution of *BRAF*^{V600E} mutations, *TERT* promoter mutations, as well as T and M stages among the clusters.

Cluster 1 exhibited the largest tumors, with a median size of 2.15 cm ($P < 0.001$), and a higher incidence of gene fusions (10.5%), although this difference was not statistically significant ($P = 0.1667$). In contrast, Cluster 3 showed the highest prevalence of *BRAF*^{V600E} mutations (92.5%, $P = 0.0099$). *TERT* promoter mutations ($P = 0.0363$) and distant metastasis (M1) ($P = 0.0326$) were more frequent in Cluster 1, although LNM did not reach statistical significance.

The American Thyroid Association (ATA) risk classification differed significantly across the clusters ($P < 0.001$), with Cluster 2 predominantly consisting of intermediate-risk patients (72.0%) and Cluster 1 containing a larger share of high-risk patients (42.1%). In addition, significant radiological feature differences were observed in calcification patterns ($P = 0.0039$) and nodule appearance under US, with heterogeneous echotexture being most prevalent in Cluster 3 ($P = 0.0085$). These findings highlight the distinct clinical and molecular characteristics of each radiomic cluster while acknowledging variables without statistical significance, supporting the potential use of this stratification approach in thyroid cancer risk assessment (Table 2).

Developing Cluster 2 specific scoring system using a machine learning algorithm

We assessed the likelihood of predicting radiomics-defined clusters using selected features. Least absolute shrinkage and selection operator (LASSO) was used to select the most effective features for scoring each cluster likelihood. A lambda value was selected to minimize binomial deviance in each cluster and facilitate the selection of radiomic features with nonzero coefficients at these optimal values for model development (fig. S1, A to F). As a result, 27, 23, and 15 radiomic features with nonzero LASSO coefficients were identified for Clusters 1 to 3, respectively. The performance of the constructed radiomic model was assessed using a test dataset. We observed an equivocal area under the curve (AUC) value of 0.98 when predicting all three clusters and ensured significant performance when compared with other known machine learning algorithms (fig. S1, G to I).

Association between radiologist interpretations and Cluster 2 score

A previous analysis identified Cluster 2 as having favorable clinicopathological features, warranting further investigation. To capture this "Cluster 2–like" signature for each patient, we developed a Cluster 2 score, a radiomics-based metric derived from a set of optimized features weighted by LASSO coefficients. This score reflects the likelihood that an individual tumor shares the defining characteristics of Cluster 2. We then applied the score to every patient in our dataset to

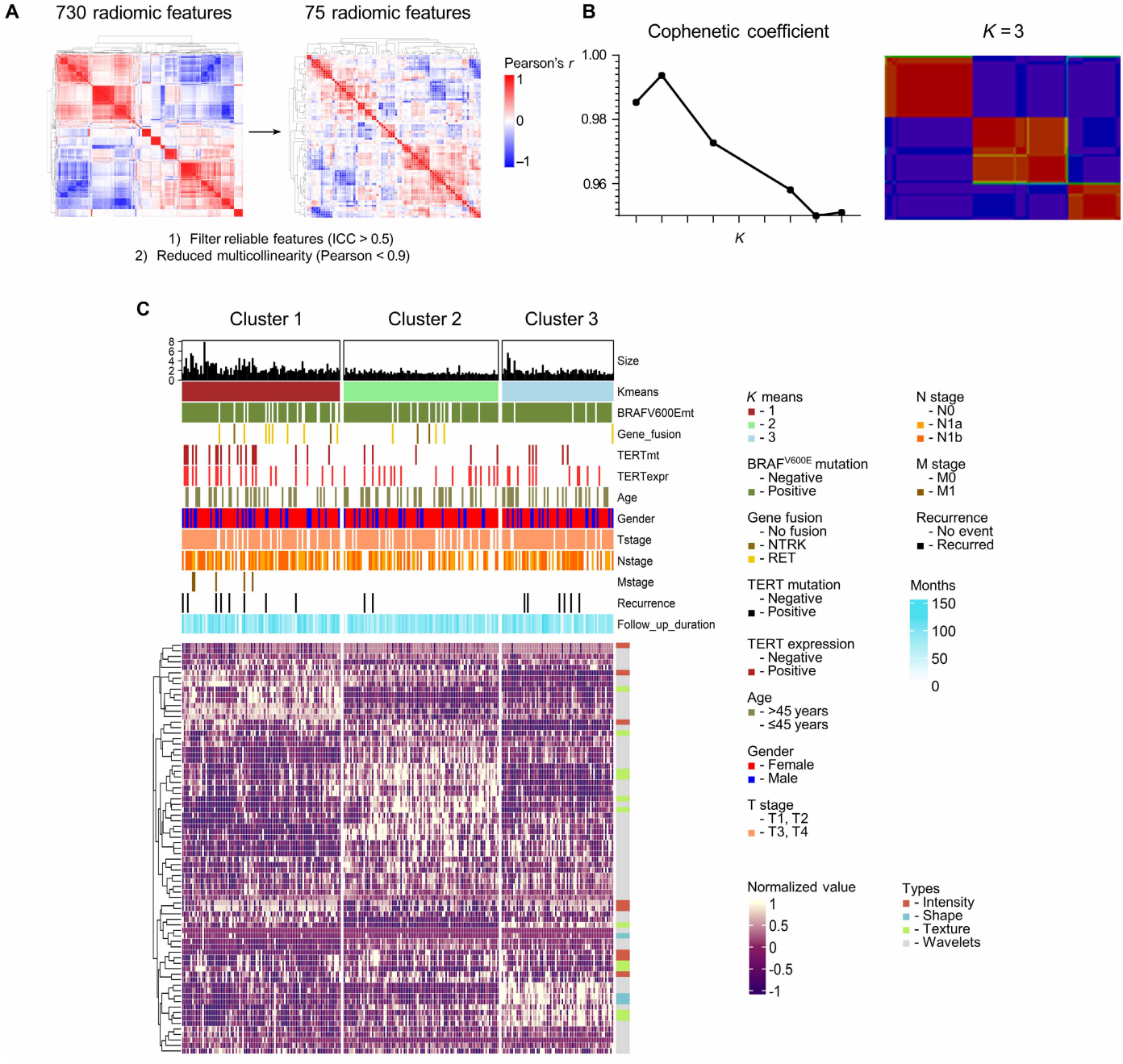


Fig. 2. Selection of reliable features and unsupervised clustering reveals three unique radiomics-based clusters. (**A**) A total of 75 of 730 radiomic features were selected after the feature selection process. A pairwise correlation plot was generated using Pearson's correlation coefficient. (**B**) A cophenetic coefficient plot and relative heatmap were drawn after unsupervised clustering using the NMF algorithm. $K = 3$ has a maximum coefficient of 0.9958, indicating that it is the optimal clustering number. (**C**) Heatmap of 75 selected features across $K = 3$ radiomic clusters with colabeled clinical variables. Hierarchical clustering was performed on the rows for better visualization.

explore its relationship with high-risk US findings as interpreted by specialized radiologists. To evaluate the clinical relevance of this score in relation to high-risk US findings identified by specialized radiologists, we designated those in the top 33% of scores as the Cluster 2 high group and those in the bottom 33% as the Cluster 2 low group from the entire dataset. The evaluated US features included composition, echogenicity, shape, margin, calcification, vascularity, nodule appearance, and final Thyroid Imaging Reporting and Data System (TI-RADS) classification (graded on a 1 to 5 scale). Among these variables, "margin" (e.g., smooth, irregular, lobulated, or extrathyroidal extension) and the overall TI-RADS grading demonstrated significant differences between the high and low Cluster 2 scoring groups. Both radiologists observed a higher proportion of smooth margins in the high-score group and more prominent extrathyroidal extension in the low-score group. In addition, radiologist 2 reported a significantly lower incidence of TI-RADS 5 in the high-score group compared to radiologist 1 (Fig. 3, A to H). Representative US images of the high- and low-score groups, along with a heatmap contrasting

Table 2. Baseline demographics and US feature comparison between proposed radiomic clusters. Significance was tested for clinical variables across defined radiomic clusters. One-way ANOVA was performed for numerical values, while the chi-square test was used for categorical variables. Statistical significance was set at $P < 0.05$. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$. ATA, American Thyroid Association; IQR, interquartile range; TI-RADS, Thyroid Imaging Reporting and Data System; TR, TI-RADS category.

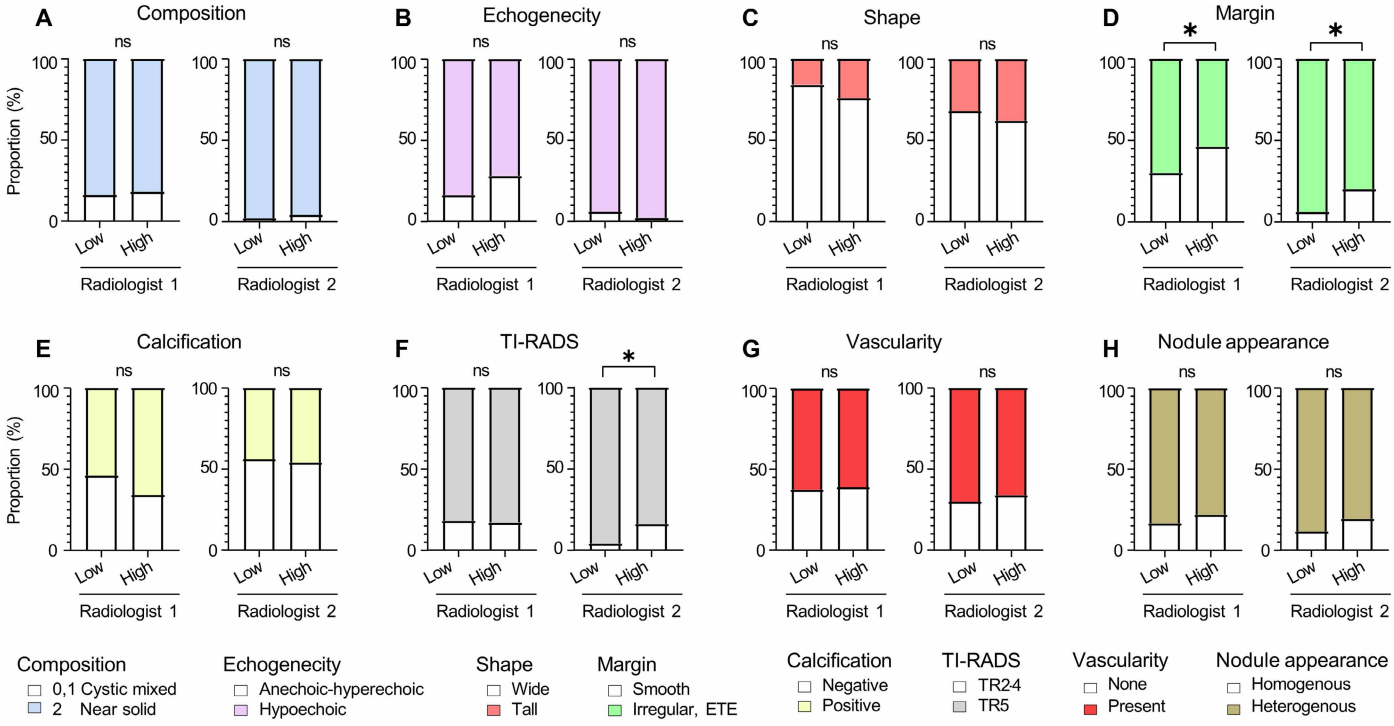
		Cluster 1 ($n = 95$)	Cluster 2 ($n = 93$)	Cluster 3 ($n = 67$)
Demographics				
Age (median, IQR)		51 (41–60.5)	52 (46–60)	52 (45–60)
Age over 45 years	No	65 (68.4%)	61 (65.6%)	42 (62.7%)
	Yes	30 (31.6%)	32 (34.4%)	25 (37.3%)
Gender	Male	27 (28.4%)	20 (21.5%)	19 (28.4%)
	Female	68 (71.6%)	73 (78.5%)	48 (71.6%)
T stage	T1	10 (10.5%)	15 (16.2%)	6 (9.0%)
	T2	7 (7.4%)	7 (7.5%)	1 (1.5%)
	T3	63 (66.3%)	64 (68.8%)	53 (79.1%)
	T4	15 (15.8%)	7 (7.5%)	7 (10.4%)
N stage	N0	24 (25.3%)	37 (39.8%)	21 (31.3%)
	N1a	30 (31.6%)	28 (30.1%)	20 (29.9%)
	N1b	41 (43.2%)	28 (30.1%)	26 (38.8%)
M stage*	M1	5 (5.3%)	0 (0%)	0 (0%)
Tumor size (cm)***		2.1 (IQR = 1.1)	1.4 (IQR = 0.4)	1.6 (IQR = 0.8)
$BRAF^{V600E}$ mutation**	Detected	78 (82.1%)	81 (87.1%)	62 (92.5%)
	Not found	85 (89.5%)	88 (94.6%)	66 (98.5%)
Gene fusion	NTRK	8 (8.4%)	3 (3.2%)	1 (1.5%)
	RET	2 (2.1%)	2 (2.2%)	0 (05)
TERT promoter mutation (C228T and C250T)*	Detected	19 (20%)	5 (5.4%)	5 (7.5%)
TERT RNA expression	Detected	20 (21.1%)	17 (18.2%)	11 (16.4%)
Recurrence	Yes	8 (8.4%)	2 (2.2%)	6 (9.0%)
Follow-up period (median, IQR)		88.5 months (IQR = 38.4 months)	86.3 months (IQR = 22.8 months)	87.7 months (IQR = 20.8 months)
ATA risk classification***	Low	15 (15.8%)	12 (12.9%)	5 (7.5%)
	Intermediate	40 (42.1%)	67 (72.0%)	34 (50.7%)
	High	40 (42.1%)	14 (15.1%)	28 (41.8%)
US features				
Composition	Cystic or mixed	12 (12.6%)	4 (4.3%)	4 (6.0%)
	Near solid	83 (87.4%)	89 (95.7%)	63 (94%)
Echogenicity	Anechoic ~ hyperechoic	85 (89.5%)	76 (81.7%)	56 (83.6%)
	Hypoechoic	10 (10.5%)	17 (18.3%)	11 (16.4%)
Margin	Smooth	56 (58.9%)	65 (69.9%)	41 (61.2%)
	Irregular, ETE	39 (41.1%)	28 (30.1%)	26 (38.8%)
Calcification**	Negative	25 (26.3%)	46 (49.5%)	23 (34.3%)
	Positive	70 (73.7%)	47 (50.5%)	44 (65.7%)
Shape	Wide	58 (61.0%)	55 (59.1%)	40 (59.7%)
	Tall	37 (39.0%)	38 (40.9%)	27(40.3%)
TI-RADS	TR 2–4	32 (33.7%)	33 (35.5%)	20 (29.9%)
	TR 5	63 (66.3%)	60 (64.5%)	47 (70.1%)
Vascularity	Negative	28 (29.5%)	41 (44.1%)	29 (43.3%)
	Positive	67 (70.5%)	52 (55.9%)	38 (56.7%)
Nodule appearance under echo**	Homogenous	38 (40.0%)	25 (26.9%)	12 (17.9%)
	Heterogenous	57 (60.0%)	68 (73.1%)	55 (82.1%)

the intensity of Cluster 2–specific radiomic features, are shown in Fig. 3 (I and J).

Cluster 2 score independently predicts advanced LNM and therapeutic indications

To assess the predictability of the Cluster 2 score for cancer stages and therapeutic indications, we performed ordinal regression analysis, controlling for clinical variables such as age, tumor size, and sex. In addition, to benchmark current clinical practice, we included the US features of nodule appearance, vascularization, and the final TI-RADS category as covariates (Table 3).

Tumor size (in centimeters) was the sole independent predictor of the T stage, with a beta estimate value of 0.301 ± 0.079 ($P = 0.001$). To predict the extent of thyroidectomy (event = total thyroidectomy), we used the Cluster 2 score, age, and tumor size as significant indicators, with beta estimate values of -0.691 ± 0.219 ($P = 0.002$), 0.217 ± 0.096 ($P = 0.025$), and 0.211 ± 0.065 ($P = 0.001$), respectively. We also assessed the ordinal prediction of Cluster 2 against the N stage, yielding a beta estimate value of -0.592 ± 0.261 ($P = 0.024$). However, the extent of neck dissection (event = lateral neck dissection) was not significantly predicted by any of these variables. In addition, we evaluated patients indicated for high-dose



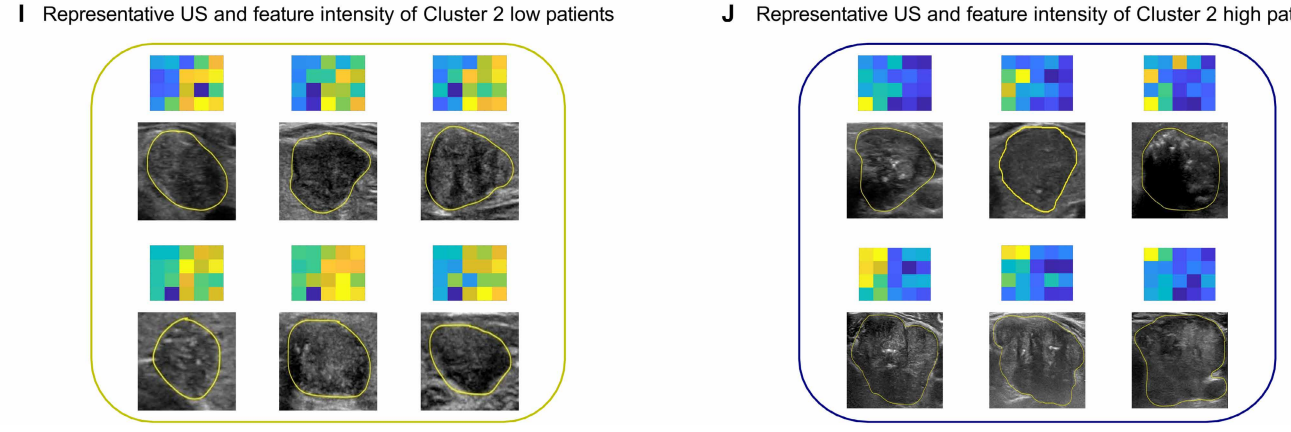


Fig. 3. Association of TI-RADS components and Cluster 2 score. Patients in the lowest and highest 33% of Cluster 2 scores were classified as low and high Cluster 2 groups, respectively. (**A** to **H**) The bar chart illustrates the proportions of each component comprising the TI-RADS and the final TI-RADS category (TR) as assessed by two specialized radiologists, as well as the proportions of vascularity and nodule appearance observed on US, with significant differences determined using the chi-square test. ETE, extrathyroidal extension. (**I** and **J**) Gross US images of tumor segmentation in Cluster 2 low and high groups are paired pixel intensities of Cluster 2 specific radiomic features found by the LASSO machine learning algorithm. The results were considered significant if the *P* value was less than 0.05. $^*P < 0.05$. ns, not significant.

radioiodine therapy and revealed that the Cluster 2 score and tumor size were significant predictors, with beta estimate values of −0.985 ± 0.508 ($P = 0.044$) and 0.587 ± 0.151 ($P < 0.001$), respectively.

In our exploratory dataset, we compared the performance regression models with and without the Cluster 2 scores using the *F* test. The predictive performance was significantly improved for thyroidectomy ($P = 0.007$) and N-stage prediction ($P = 0.036$) when the Cluster 2 score was included.

To validate this regression model, we assessed whether the inclusion of the Cluster 2 score significantly enhanced its predictive performance. Specifically, for T3 and T4 cancer stages, the inclusion of the Cluster 2 score did not improve predictability (Fig. 4A). However, its inclusion significantly enhanced the prediction of total thyroidectomy events, as reflected by AUC values of 0.768 ($P = 0.024$) in the internal validation dataset and 0.786 ($P = 0.042$) in the external dataset (Fig. 4B). Similarly, the Cluster 2 score significantly improved the prediction of lateral neck node metastasis (N1b stage), with AUC values of 0.781 ($P = 0.030$) internally and 0.808 ($P = 0.005$) externally (Fig. 4C). Although the prediction of distant pathological LNM improved, the Cluster 2 score did not significantly enhance the prediction of lateral neck dissection (Fig. 4D). Regarding adjuvant radioiodine therapy, the inclusion of the Cluster 2 score significantly improved the model's ability to predict clinician decisions for identifying patients requiring high-dose therapy, with AUC values of 0.788 ($P = 0.010$) and 0.750 ($P = 0.006$) for the internal and external validation datasets, respectively (Fig. 4E).

Cluster 2 score predicts poor outcomes and aggressive PTC features

We further investigated the prognostic significance of the Cluster 2 score in patients with PTC. First, we evaluated the 1-year thyroglobulin (Tg) level as a reliable biomarker for assessing treatment response, recurrence risk, and residual disease. We selected patients who underwent total thyroidectomy and collected their nonstimulated Tg from 1-year follow-up laboratory results. The detection cutoff was set at 1 μg/liter, and post–total thyroidectomy patients with Tg levels above this threshold were considered to have a biochemical incomplete response. We observed that patients with detectable Tg levels showed lower Cluster 2 scores in both exploratory and internal validations. However, this finding was not replicated in the external validation dataset, although a similar trend was observed (Fig. 5A).

Considering that driver mutations are major indicators of tumor aggressiveness in PTC, we compared the representative mutations with their corresponding Cluster 2 scores (Fig. 5B). *BRAF*^{V600E} mutation alone was not significantly different from the wild type. However, its coexistence with the *TERT* promoter mutation was significant across all datasets.

Our exploratory dataset had an average follow-up period of 7.2 years, allowing us to evaluate disease-free survival (DFS) in relation to the Cluster 2 score. Patients with high Cluster 2 scores demonstrated significantly longer DFS than those with low Cluster 2 scores (Fig. 5C). Moreover, we were interested in benchmarking the predictive use of the Cluster 2 score with the established ATA risk stratification system; therefore, we conducted a multivariable Cox proportional hazards regression analysis incorporating both variables. In this model, the ATA risk classification remained a strong predictor of DFS [hazard ratio (HR) = 2.11, 95% confidence interval (CI): 0.92 to 4.83, $P = 0.0784$], while higher Cluster 2 scores demonstrated a trend toward reduced recurrence risk (HR = 0.17, 95%

Table 3. Ordinal logistic regression analysis of radiomics-specific cluster scores and clinical factors in exploratory patient datasets. Ordinal logistic regression analysis was conducted for the following variables: T stage (T1–2 = 0, T3–4 = 1), extent of thyroidectomy (partial thyroidectomy and lobectomy = 0, total thyroidectomy = 1), N stage (N0 = 0, N1a = 1, N1b = 2), extent of neck dissection (central neck node dissection = 0, modified lateral neck node dissection = 1), and dosage of adjuvant radioiodine therapy (none = 0, low dose <60 mCi = 1, intermediate dose 100 to 120 mCi = 2, high dose >150 mCi = 3). An *F* test was conducted to compare the predictive performance of the two models: one model used clinical factors, including demographic features and radiological interpretations, while the other model incorporated radiomic cluster scores in addition to clinical factors. Significant results from the *F* test indicated that predictive performance improved with the addition of radiomic cluster scores. The estimates were considered significant if the *P* value was <0.05. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

	Exploratory dataset	
	Estimate	**Standard error**
T stage		
Cluster 2 (score)	−0.083	0.269
Age (year)	0.060	0.117
Tumor size (cm)***	0.301	0.079
Sex	0.017	0.131
Nodule appearance	−0.001	0.003
Vascularity (%)	−0.006	0.003
TI-RADS category	0.186	0.158
F test	$P = 0.631$	
Extent of thyroidectomy	**Exploratory dataset**	
	Estimate	**Standard error**
Cluster 2 (score)**	3.829	0.725
Age (year)*	−0.691	0.219
Tumor size (cm)**	0.217	0.096
Sex	0.211	0.065
Nodule appearance	0.092	0.107
Vascularity (%)	0.001	0.002
TI-RADS category*	−0.001	0.003
F test	$P = 0.007$	
N stage	**Exploratory dataset**	
	Estimate	**Standard error**
Cluster 2 (score)*	−0.592	0.261
Age (year)	0.093	0.114
Tumor size (cm)	0.099	0.077
Sex	−0.211	0.127
Nodule appearance	−0.001	0.003
Vascularity (%)	0.002	0.004
TI-RADS category	0.188	0.153
F test*	$P = 0.036$	
Extent of neck dissection	**Exploratory dataset**	
	Estimate	**Standard error**
Cluster 2 (score)	−0.321	0.182
Age (year)	−0.026	0.080
Tumor size (cm)	0.065	0.054
Sex	−0.063	0.089

(*Continued*)

(*Continued*)

	Exploratory dataset	
	Estimate	**Standard error**
Nodule appearance	−0.001	0.002
Vascularity (%)	0.002	0.003
TI-RADS category	−0.001	0.107
F test	$P = 0.327$	
RAITx dosage	**Exploratory dataset**	
	Estimate	**Standard error**
Cluster 2 (score)*	−0.985	0.508
Age (year)	−0.051	0.223
Tumor size (cm)***	0.587	0.151
Sex	−0.298	0.248
Nodule appearance	0.000	0.006
Vascularity (%)	0.004	0.008
TI-RADS category	−0.171	0.299
F test	$P = 0.089$	

CI: 0.02 to 1.37, $P = 0.0947$), although this did not reach statistical significance. Nonetheless, the combined model significantly enhanced predictive performance ($P = 0.007$, concordance = 0.708), suggesting that integrating radiomics-based stratification with traditional ATA risk levels may improve clinical risk assessment (data S2).

In addition, histological variants, which also indicate tumor aggressiveness, showed significantly low Cluster 2 scores in the presence of rare but aggressive histological types (Fig. 5D). A summary of the aggressive histological variants found in our datasets is presented in a pie chart in Fig. 5E. In the exploratory dataset, we identified 11 cases: five solid variants, four diffuse sclerosing variants, one tall cell variant, and one oncocytic variant. In the internal validation set, we found nine cases: three diffuse sclerosing variants, five tall cell variants, and one oncocytic variant. Meanwhile, we found 12 cases in the external validation set: two solid variants, five diffuse sclerosing variants, and five tall cell variants.

Integration of transcriptomics analysis identifies distinct biology in Cluster 2

To explain the potential of our scoring system for predicting clinical outcomes in patients with PTC, we incorporated molecular interpretations. We defined the biological characteristics of the three radiomic clusters using gene set enrichment analysis (GSEA) of 51 hallmark gene sets known to contribute to tumorigenesis (*29*). A heatmap was plotted showing the average normalized enrichment scores of the significant gene sets (Fig. 6A). Specifically, Cluster 2 was up-regulated in PANCREAS_BETA_CELLS, ANDROGEN_RESPONSE, ESTROGEN_RESPONSE_EARLY, and GLYCOLYSIS. Cluster 3 was up-regulated in UNFOLDED_PROTEIN_RESPONSE, UV_RESPONSE, APICAL_JUNCTION, INTERFERON_ALPHA_GAMMA_RESPONSE, COMPLEMENT, PEROXISOME, BILE_ACID_METABOLISM, ALLOGRAFT_REJECTION, EPITHELIAL_MESYNCHYMAL_TRANSITION, and INFLAMMATORY_RESPONSE groups. Cluster 1 was down-regulated in all significant hallmark pathways.

Subsequently, we benchmarked well-known transcriptomic markers to analyze the aggressiveness of PTC. The thyroid differentiation score, where a higher value indicates better differentiation, was significant when directly comparing Clusters 2 and 3 (Fig. 6B). However, no significant differences were found when comparing the highest (top 33%) and lowest (bottom 33%) Cluster 2 scores (Fig. 6C). BRAF and RAS, two major driver mutations in thyroid cancer, tend to have opposing effects on tumorigenesis, with BRAF being more aggressive than RAS. The BRAF-RAS score was higher in Cluster 3, indicating a greater likelihood of *BRAF* mutations (Fig. 6D). Moreover, the Cluster 2 high group exhibited a significantly higher RAS-like profile than the Cluster 2 low group (Fig. 6E).

We delineated the distinct biology of Cluster 2 by analyzing the differentially expressed genes (DEGs). DEGs with an adjusted $P < 0.05$ ($\log_{10} P$ value =1.301) and an absolute $\log_2$ fold change of >0.5 were considered significant. This analysis returned 306 significant DEGs in the Cluster 2 high group and 198 DEGs in the Cluster 2 low group (Fig. 6F and data S3). We subsequently profiled these DEGs using the Gene Ontology database, revealing that the Cluster 2 low group was associated with mitochondrial respiration and lipid metabolism (Fig. 6G and data S4), whereas the Cluster 2 high group was enriched with gene sets involved in immune response, lymphocyte activation, immunoglobulin complexes, and cell differentiation (Fig. 6H and data S5).

Among the top 20 DEGs found in both the Cluster 2 high and low groups, we validated five representative DEGs from each cluster in a validation dataset using qPCR (Fig. 7, A to J). Validated DEGs were significantly associated with prognosis. The expression of Cluster 2 high–specific DEGs correlated with better DFS (Fig. 7K), whereas higher expression of Cluster 2 low DEGs was associated with worse DFS (Fig. 7L). Likewise, we were able to validate the uniform prognostic value of Cluster 2–defined DEGs in The Cancer Genome Atlas (TCGA) thyroid cancer patient dataset (fig. S2, A and B).

From the validated DEGs, we selected paired box 5 (PAX5), a B cell transcription factor, and activating transcription factor 5 (ATF5), which is known to mediate mitochondrial unfolded protein responses, for further protein expression validation (*30*, *31*). Fresh frozen paraffin slides from surgical specimens of the Cluster 2 high ($n = 15$) and low ($n = 15$) groups were used for this validation (fig. S2, C and D). IHC was conducted to evaluate the expression levels of PAX5 and ATF5, which revealed significant differences in the proportion of positively stained cells between the groups (Fig. 7, M and N).

For PAX5 expression, the Cluster 2 low group had 0.2% high positive, 0.4% moderate positive, 5.3% low positive, and 94.1% negative expression, while the Cluster 2 high group exhibited 1.4% high positive, 5.6% moderate positive, 17.2% low positive, and 75.8% negative expression (Fig. 7O). For ATF5 expression in the Cluster 2 low group, 2.9% were highly positive, 20.8% moderately positive, 40.7% low positive, and 35.6% negative. In contrast, the Cluster 2 high group showed 0.1% positive, 4.1% moderate positive, 25.6% low positive, and 70.2% negative expression (Fig. 7P). In addition, staining of normal thyroid tissues for ATF5 and PAX5 showed no detectable expression, suggesting that their expression was likely tumor specific (fig. S2E).

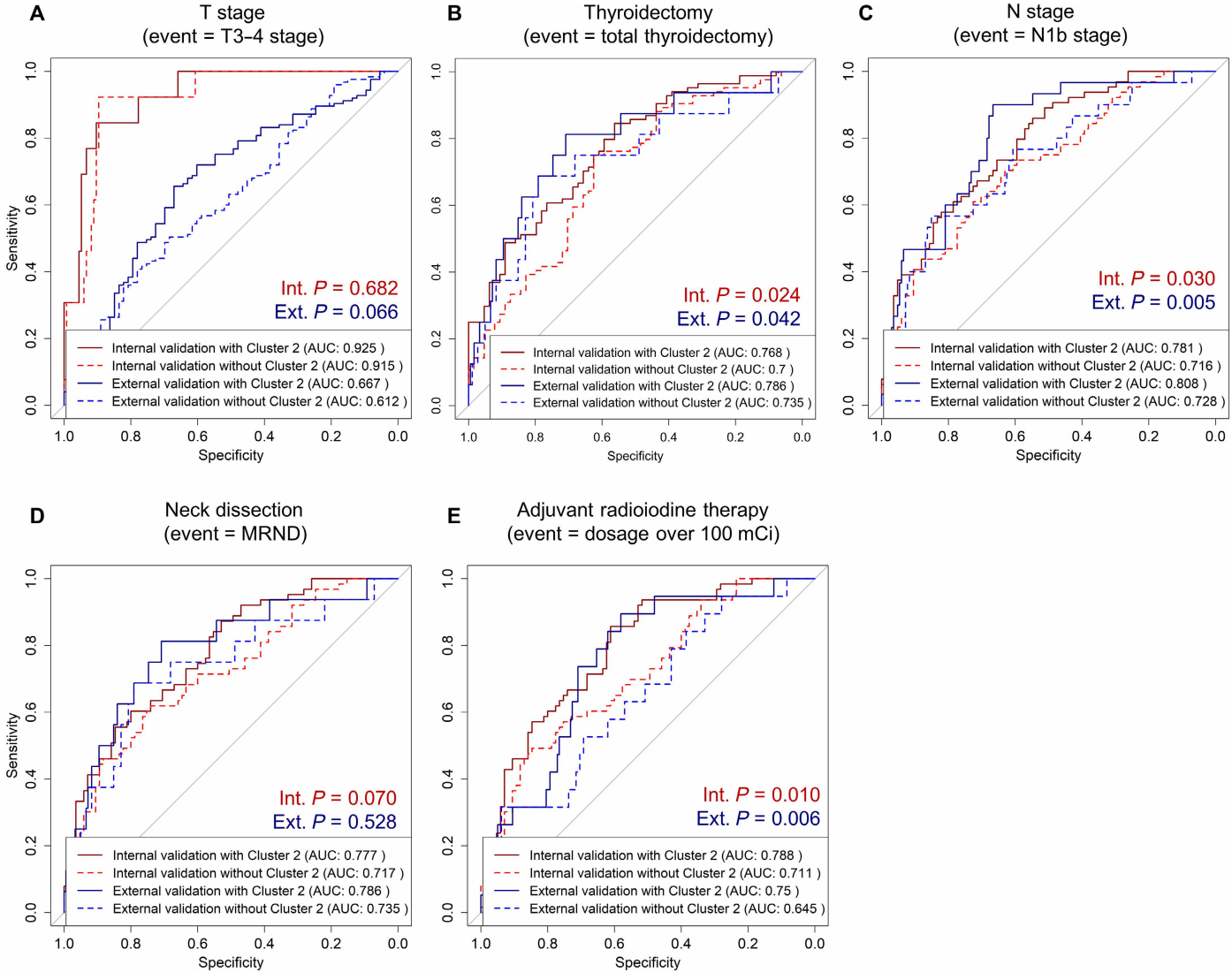


Fig. 4. Validation of the prognostic value of Cluster 2 score using ordinal logistic regression model. Predictions were assessed using receiver operating characteristic (ROC) curves for (**A**) T stage (event: T3–4 stage), (**B**) extent of thyroidectomy (event: total thyroidectomy), (**C**) extent of neck dissection [event: modified lateral neck node metastasis (MRND)], (**D**) N stage (event = N1b stage), and (**E**) indication of moderate radioiodine therapy (event: radioiodine therapy >100 mCi). The regression model from the internal validation patient dataset is depicted in red, whereas the external validation dataset is depicted in blue. The dotted line represents the model performance with only clinicopathological variables, and the solid line represents the model with the inclusion of the Cluster 2 score in both the internal and external models. The significance of the improved model performance with the inclusion of Cluster 2 was assessed using the DeLong test, and the results are shown in the legend (analysis between solid and dotted lines). Results were considered significant if the *P* value was lower than 0.05.

DISCUSSION

To date, various studies have used radiomics to identify thyroid US imaging biomarkers as predictors of thyroid cancer outcome or prognosis. However, most of these studies have primarily focused on distinguishing malignant tumors from normal tissue or detecting possible LNM (*32*). Radiomics studies on thyroid cancer typically targeted a known, single-aggressive clinical factor as the end point of the study; however, no satisfactory model has been proposed. In the initial stage of the study, we aimed to enhance our investigation by enrolling only highly correlated radiomic features with clinical factors. However, this approach returned comparatively poor clustering performance, and the characterization of formed clusters was unsatisfactory (fig. S3).

For genomic markers, the performance of the previous radiomic prediction models was not significant, with an average AUC value of ~0.7 (*33*, *34*). Similarly, we failed to predict the $BRAF^{V600E}$ mutation alone. However, recent bioinformatics analysis revealed a distinct behavior for the $BRAF^{V600E}$ mutation in PTC, which may function as a lurking variable for prediction models (*24*). However, the coexistence of $BRAF^{V600E}$ and *TERT* promoter mutations defines the most aggressive form of PTC, with distinct biological characteristics, which our scoring system successfully identified (*16*, *35*, *36*).

Predicting malignancy is not a pressing need at the screening US examination stage, as specialized radiologists effectively fulfill their roles (*37*). Particularly, current radiological algorithms, such as TI-RADS, achieve sensitivities of >90% in diagnosing thyroid nodules.

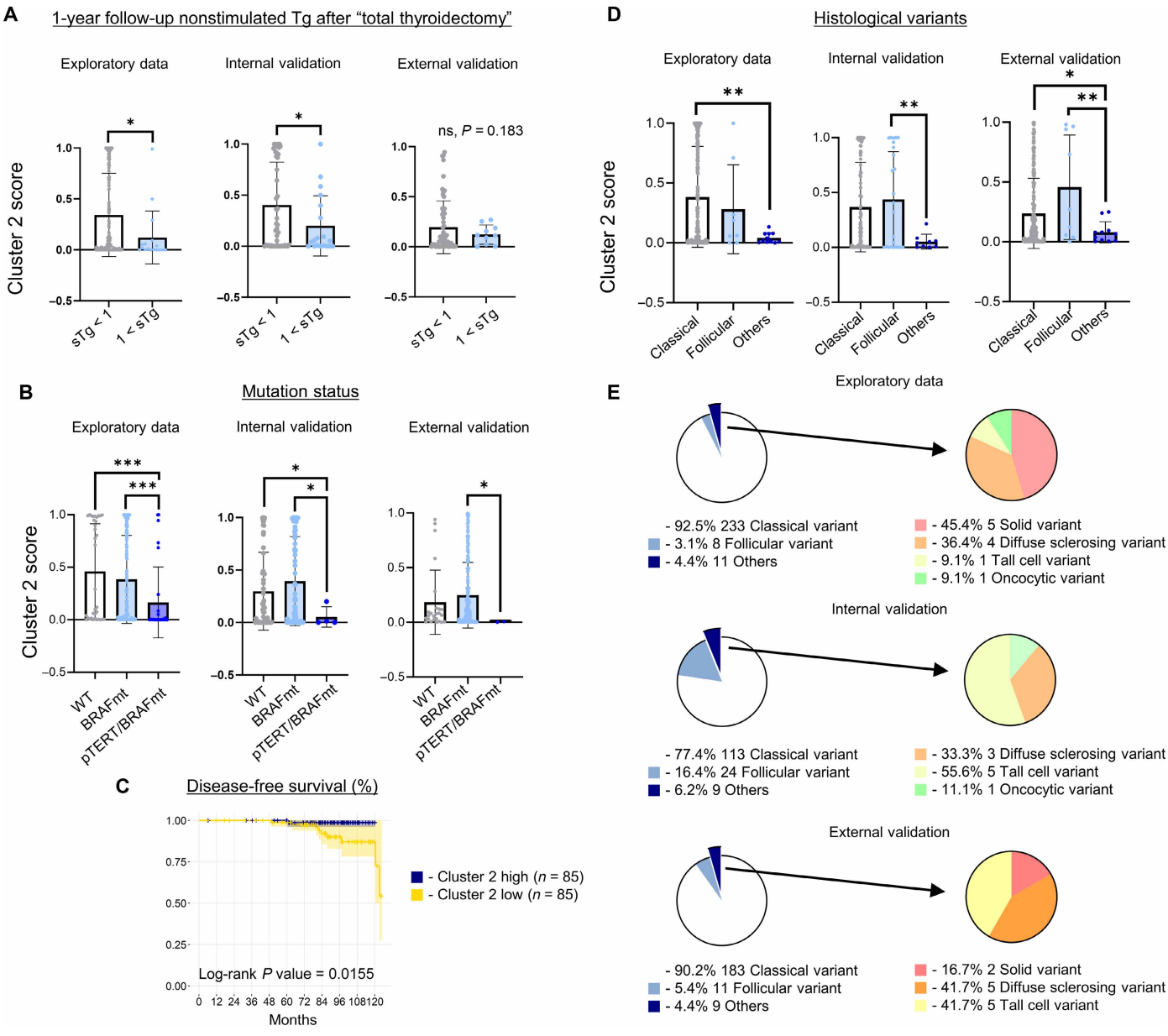


Fig. 5. Association of Cluster 2 score with response to therapy and aggressive biological features. Measurements were conducted separately for each dataset. (**A**) Comparison of Cluster 2 scores between patients with nonstimulated Tg levels in 1-year follow-up laboratory results greater than 1 and those without. sTg, stimulated thyroglobulin. (**B**) Comparison of Cluster 2 scores between patients who were intact from the mutation, those with the *BRAF*^{V600E} mutation (*BRAF*mt), and those with both *TERT* promoter (pTERT) mutation and *BRAF*^{V600E} mutation. WT, wild type. (**C**) The Kaplan-Meier survival curve of disease-free survival (DFS) was generated based on the Cluster 2 score, with high and low groups defined by the 33rd and 66th percentiles, respectively. Statistical significance was evaluated using the log-rank test. (**D**) Comparison of Cluster 2 scores between patients diagnosed with classical variant PTC, follicular variant PTC, and other aggressive histology of PTC. (**E**) The layout of the number and types of aggressive histologies in each dataset is presented in a pie chart. The Mann-Whitney *U* test was used to discern the significance of numerical values. Results were considered significant if the *P* value was lower than 0.05. Every bar chart in this figure represents the average values, with error bars indicating standard deviations. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

However, this comes with a trade-off of lower specificity, which can lead to unnecessary concerns about overtreatment (*38*, *39*). In addition, visual analysis by physicians is highly subjective, leading to both interobserver and intraobserver variations (*40*, *41*). In contrast, radiomics serves as an objective imaging biomarker by extracting quantitative features from images, providing insights into underlying pathophysiology that are impossible to discern through visual interpretation alone (*42*). Although US–fine-needle aspiration or biopsy helps determine the need for surgery, the increasing use of high-quality imaging methods facilitates the detection of even small thyroid cancers, thereby increasing the risk of overtreatment. This has prompted the application of AS in recent years (*43*–*45*). However, it remains unclear which cancers may progress or have poor outcomes during AS. To address this gap, we characterized radiomic

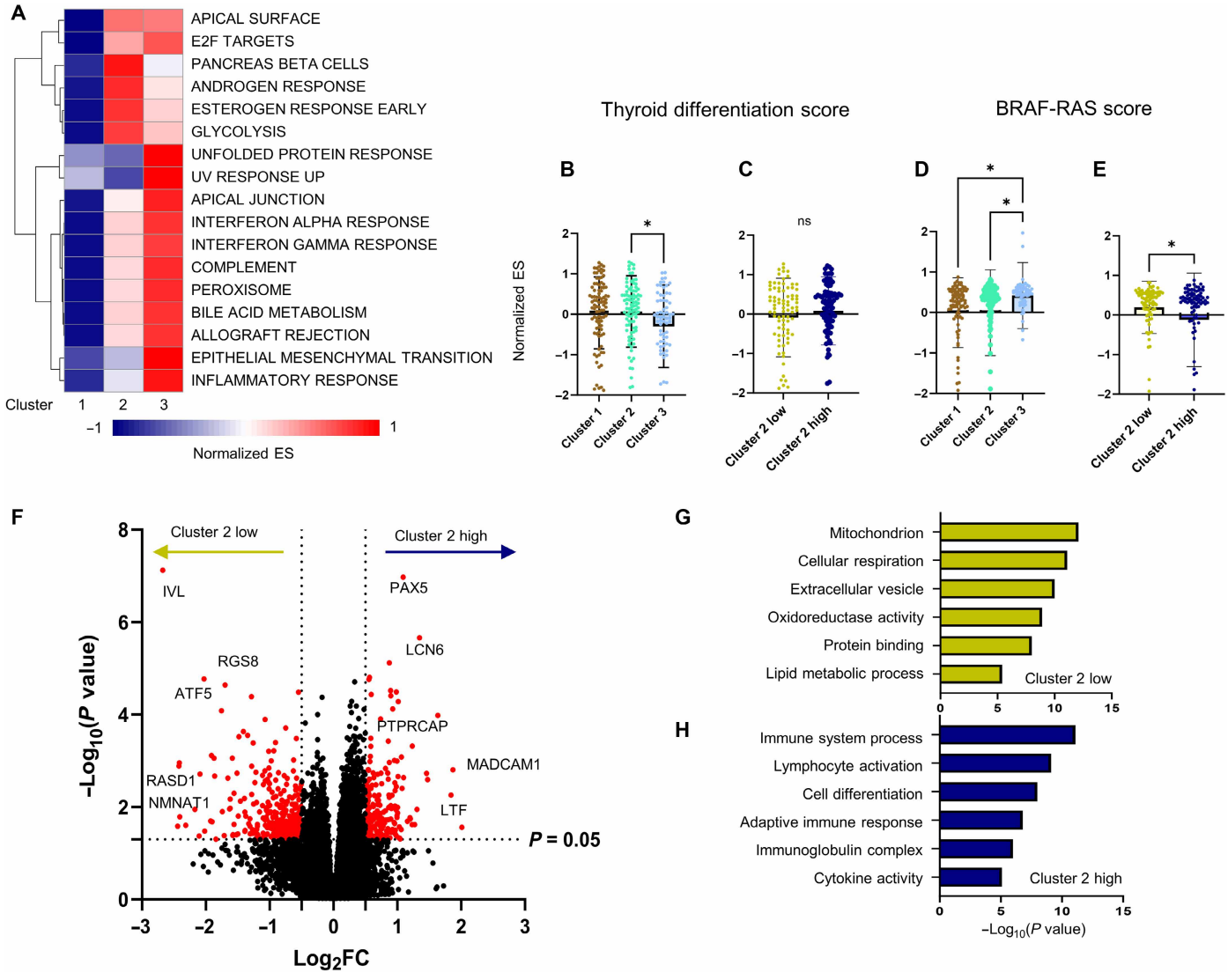


Fig. 6. Molecular characterization of radiomic clusters. (**A**) A heatmap of enrichment scores (ES) retrieved after conducting GSEA in radiomic clusters using hallmark gene sets of tumorigeneses registered in the molecular signature database (mSigDB). Only gene sets with significant differences in one-way ANOVA analysis are shown. (**B** and **D**) Thyroid differentiation scores and (**C** and **E**) BRAF-RAS scores were calculated using single-sample GSEA and compared across the three radiomic clusters, as well as between Cluster 2 high and low patient subgroups, defined by the top and bottom 33% of the Cluster 2 score distribution, respectively. (**F**) Significant differentially expressed genes (DEGs) found between Cluster 2 high and low patient groups in the exploratory dataset are depicted as volcano plots. (**G** and **H**) DEGs from each Cluster 2 high and low patient group were profiled using gene ontology gene sets, and the top six significant gene sets are shown in bar charts. The Mann-Whitney *U* test was used to discern the significance of numerical values. Results were considered significant if the *P* value was lower than 0.05. All bar charts are depicted as averages, and error bars indicate standard deviations. *$P < 0.05$.

clusters, a biomarker that predicts PTC with favorable characteristics, which could provide notable improvement in clinical practice by reducing unnecessary treatments.

Our scores successfully predicted the pathological N stage of cancer, although their performance was suboptimal for predicting the extent of neck dissection required. These findings suggest a potential role for radiomics screening in aiding physicians in surgical decision-making. We observed a significant difference in tumor size among radiomic clusters and hypothesized that tumor size could be a primary determinant of Cluster 2 score. However, its predictive role in the T stage was limited when tumor size was analyzed as a covariate, suggesting that Cluster 2 was influenced by different, unmeasured radiological factors indicative of indolence.

The "black box issue" is an inherent limitation of radiomics methodology as it lacks transparency and is not readily interpretable by clinicians (*46*). This issue is prevalent among AI techniques applied in medicine and is often scrutinized because of the critical need for understanding before these methods can be confidently used in clinical decisions (*47*).

To address these limitations, we integrated a multiomics approach to gain biological insights into the association between our Cluster 2 score and clinical outcomes. GSEA results using hallmark gene sets revealed enrichment of inflammatory and epithelial-to-mesenchymal transition (EMT) signals in Cluster 3. This aligns with our previous findings, where we identified a significant association between the consistent up-regulation of immune response and EMT signals in suspicious nodules identified by radiologists, features that

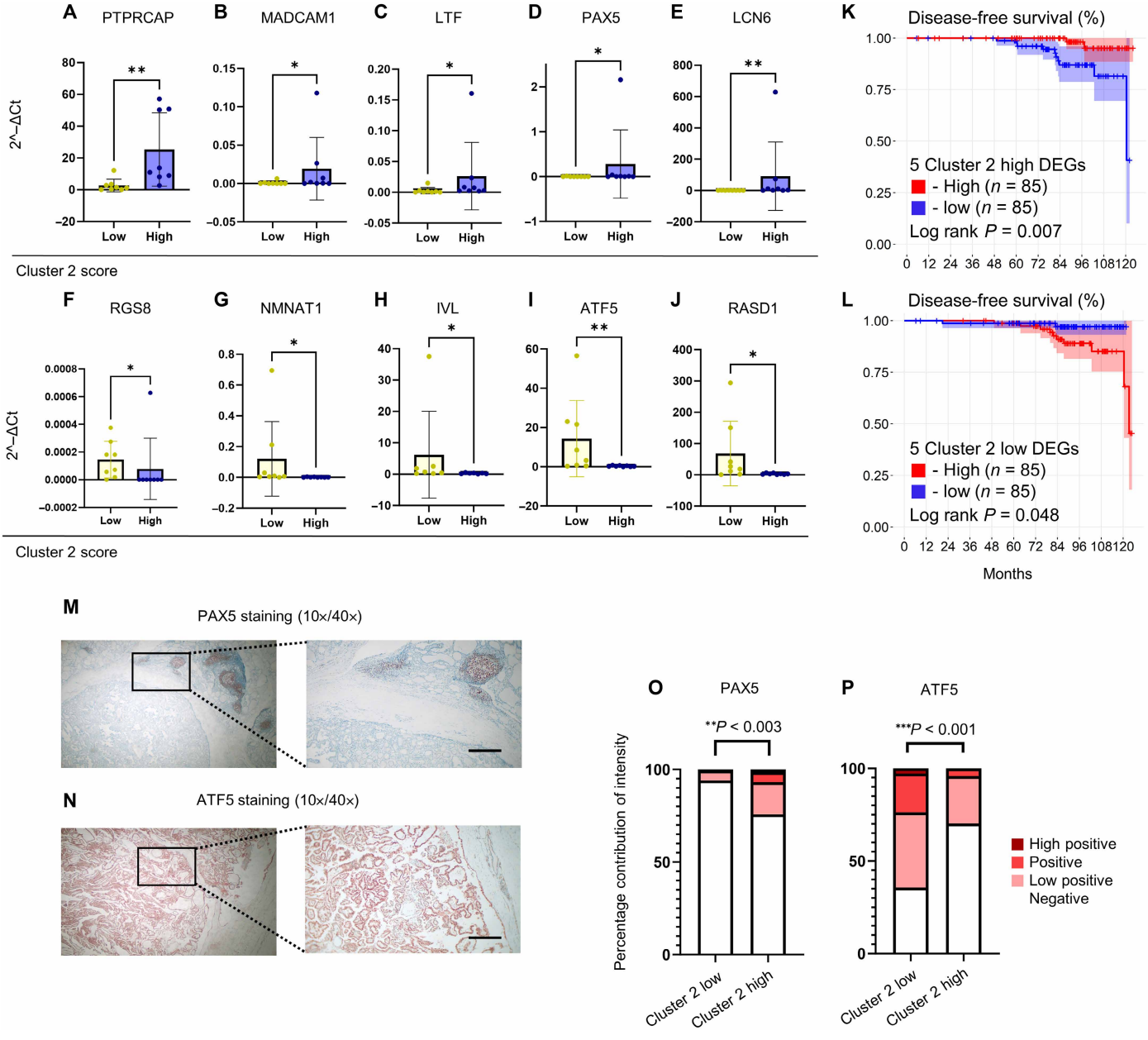


Fig. 7. Validation of molecular features correlated with Cluster 2 scores. Among the DEGs with the highest significance, five DEGs from each (**A** to **E**) Cluster 2 high and (**F** to **J**) Cluster 2 low were validated in our patient tissue samples using qPCR ($n = 7$ tissue samples for each Cluster 2 high and low patient group). Kaplan-Meier survival curves for DFS were plotted using validated DEG markers from our data. The association of five Cluster 2 high–specific DEGs with DFS is depicted in (**K**), while the association of five Cluster 2 low–specific DEGs with DFS is depicted in (**L**). The log-rank test was used to assess statistical significance. (**M** and **N**) Representative images from IHC staining for PAX5 and ATF5 are shown in 10× and 40× (scale bar in 40× represent 200 μm), (**O** and **P**) and the percent contribution of pixels with differing staining intensity analyzed via ImageJ was compared between the Cluster 2 high and low groups. The definitions of the Cluster 2 low and high signature groups were determined by 33 and 66% cutoffs, respectively. The Mann-Whitney *U* test was applied to discern the significance of numerical values. Results were considered significant if the *P* value was lower than 0.05. All bar charts are depicted as averages, and error bars indicate standard deviations. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

closely resemble those of Cluster 3 identified in the current study. (*48*). In Cluster 2, an increase in different hormonal responses and gene set enrichment, which are important for cell development, was observed. In contrast, Cluster 1 showed no significantly enriched gene sets, indicating a "silent biology." We interpret this lack of transcriptional activity as reflective of a heterogeneous PTC population. Although the overall cohort was genomically homogeneous, with ~85% of patients harboring the $BRAF^{V600E}$ mutation, Cluster 1 exhibited a comparatively higher proportion of fusion genes and TERT promoter mutations. This subtle molecular heterogeneity highlights the need for further subclassification within this group.

Summarizing our overall genomic and transcriptomic analyses, our radiomic clusters can be characterized as follows: Cluster 1 is associated with a silent but heterogeneous genetic background, Cluster

2 exhibits an RAS-like profile, and Cluster 3 exhibits a BRAF-like profile. This classification aligns with the genomic characterization of PTC as defined in a TCGA study (*11*).

We also explored whether differences in cell proportions could be detected using radiomic features. To this end, we calculated the immune, stromal, and ESTIMATE scores using the xCell algorithm. However, we did not find any significant associations, suggesting that the infiltration of immune cells, cells of mesenchymal origin, and tumor purity may not affect our scoring model (fig. S4, A to C). Nonetheless, specific immune cell types were differentially infiltrated with respect to the Cluster 2 score, indicating that distinct immune responses may drive each tumor condition (fig. S4, D to I).

Recent epidemiological data indicate that microcarcinomas may comprise over half of newly diagnosed PTC, underscoring the importance of improved risk stratification in lower-risk populations (*49*). By validating our model in demographically distinct cohorts, we confirmed its performance across a broader clinical spectrum. This highlights the potential of the Cluster 2 score to guide more personalized treatment decisions, including AS. However, additional multicenter validation is necessary, particularly considering the relevance of *BRAF* mutation status and other molecular markers. We plan to conduct future prospective studies in collaboration with various other institutions. However, most medical institutions adhere to their own formats to save radiological databases, which complicates data sharing (*50*). Therefore, establishing a universal guideline for database storage is crucial to facilitate seamless data communication. Furthermore, benchmarking strategies and combinatorial studies involving deep-learning algorithms in conjunction with our radiomics approach are warranted. These tools are pivotal in the emerging AI-medicine paradigm and can enhance the accuracy and applicability of our model. Establishing standardized practices and incorporating advanced AI techniques are essential for advancing the integration of radiomics into routine clinical use.

We acknowledge a limitation of this study due to the exclusive use of the American College of Radiology (ACR) TI-RADS. The reason for selecting ACR-TI-RADS was based on studies showing a correlation between the US phenotype and prognosis, where fewer concerning appearances were associated with better outcomes (*51*, *52*). Although TI-RADS is widely used in the United States, a global consensus on thyroid imaging reporting systems is under consideration, with variations such as the EU TI-RADS (European Thyroid Imaging Reporting and Data System), K TI-RADS (Korean Thyroid Imaging Reporting and Data System), and C TI-RADS (Chinese Thyroid Imaging Reporting and Data System) (*53*). Addressing these regional differences is an important perspective for future research. In addition, we acknowledge the potential risk of overfitting in our model owing to the relatively small sample size and the inclusion of a large number of radiomic features. To mitigate this, we used rigorous feature selection techniques, incorporated multiple validations with external datasets, and integrated the clinical and molecular interpretations. Nonetheless, further evaluation with larger and more diverse cohorts is required to ensure broader applicability and robustness.

Here, we refined the radiomic features by reducing and selecting 75 robust features to address the complexity issues inherent in radiomics. As a result, we identified three distinct radiomic profiles, each resembling distinct PTC molecular subtypes, which offer a topologically advantageous framework for predicting tumor characteristics, particularly in discerning indolent tumors that may be suitable for AS, an area often underexplored in conventional imaging assessments. Our radiomic Cluster 2 scoring model is designed to complement, rather than replicate, TI-RADS predictions, providing an additional, data-driven metric to support nuanced clinical decision-making. This approach highlights the potential of radiomics to refine patient selection for AS, thereby reducing overtreatment in cases where the 5-year overall survival exceeds 95%. Ultimately, this study provides a perspective on the prognostic effect of radiomics in general oncology and calls for further prospective research and validation with a larger patient cohort, incorporating multicenter trials for practical application.

MATERIALS AND METHODS

Patient enrollment

A total of 255 patients with PTC who underwent thyroidectomy at the Yonsei Cancer Center (Seoul, South Korea) between May 2014 and January 2018 were enrolled. For internal validation, we enrolled an independent cohort of 150 patients with PTC, including those diagnosed with microcarcinoma, who underwent thyroidectomy between January 2018 and January 2023. External validation included 203 patients with PTC who underwent thyroidectomy at the Yongin Severance Hospital (Yongin-si, Gyeonggi-do, South Korea) between January 2018 and January 2023. All tissue samples were snap-frozen in liquid nitrogen immediately after surgical removal and stored at −80°C until further use. Prior to surgery, all of the included patients had cytology classified as Bethesda V or VI. Risk stratification of patients regarding ATA was done using the 2015 ATA Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer, which provide evidence-based criteria for categorizing patients into low, intermediate, and high-risk groups based on clinicopathologic features (*54*).

US imaging and assessment by radiologists

At our institution, all patients undergo preoperative staging US before thyroid surgery. During the study period, these examinations were performed by one of 23 radiologists. This group comprised five radiologist staff members with 3 to 25 years of experience and 18 fellows with 1 to 2 years of experience, all specializing in thyroid imaging. High-frequency linear transducers (5 to 12 MHz) (iU22 or EPIQ 5, Philips Healthcare, Bothell, WA, USA) were used. During staging examinations, individual US features of cancers were prospectively analyzed and recorded in our institutional database (*55*, *56*).

Thyroid nodules were classified as solid, predominantly solid (cystic portion <50%), or predominantly cystic (cystic portion ≥50%). Echogenicity was classified as hyperechoic, isoechoic, hypoechoic (compared with the surrounding thyroid parenchyma), or markedly hypoechoic (compared with the adjacent strap muscle). Margins were categorized as circumscribed or noncircumscribed (microlobulated or irregular). Calcifications were classified as absent, macro- or eggshell, and micro- or mixed calcifications. Shape was classified as parallel or nonparallel, with parallel shapes being taller-than-wide, where the anteroposterior dimension exceeded the transverse dimension. Echotexture of the thyroid parenchyma was assessed as homogeneous or heterogeneous (coarse-appearing echotexture, marginal nodularity, increased/decreased anteroposterior diameter of the gland, or increased/decreased parenchymal echogenicity) (*57*). On the basis of the individual US features, two staff radiologists (J.H.Y. and J.Y.K.) with 15 and 27 years of experience in thyroid imaging, respectively,

independently provided assessments according to the ACR TI-RADS (*58*). Vascularity was assessed using two-dimensional Doppler scans and categorized into three patterns: reduced or absent, indicating the absence of Doppler signals within the thyroid nodule; peritumoral, indicating the presence of Doppler signals around the periphery of the nodule; and intratumoral, indicating the presence of Doppler signals within the thyroid nodule despite peripheral vascularity.

High-throughput radiomics feature extraction and feature selection

Prior to radiomics extraction, the representative US image of the PTC mass was selected by the radiologist (J.Y.K.) in this study. A polygonal region of interest was drawn along the border of the PTC using US images selected by the two radiologists.

Radiomics is a technique in which quantitative features and characteristics are extracted from medical images. Medical images can exhibit variability in terms of intensity values owing to differences in imaging devices, institutions, and clinical settings, which affects reproducibility. To ensure consistency of intensity values in images obtained across various settings, we use min-max normalization, which not only standardizes the images but also enhances the performance of the machine learning processes by improving pattern recognition and estimation accuracy. Once the images are normalized, pixel-intensity distribution–based features (e.g., entropy, energy, kurtosis, skewness, and median) and texture features (derived from gray-level co-occurrence and gray-level run-length matrices) are collected. In addition, wavelet transformation with the Coiflet family was used to decompose the image into low-to high-frequency modes along the *x* and *y* directions, and the same features were subsequently gathered from the resulting subimages. As a result, 730 features were returned for downstream analysis

To ensure the reliability of the measured features, we computed the ICC for each radiomic feature by changing wavelets using the "irr" R package. Two-way random effects with a single measurement and absolute agreement were applied. Given that our average ICC value was ~0.4 (Q1 = 0.058 and Q3 = 0.537), we considered features with an ICC ≥0.5 to be optimally reliable (fig. S5). To improve the efficacy of unsupervised clustering, we reduced multicollinearity by removing radiomic features with high correlations with other features. Pearson correlation analysis was performed for all candidate radiomic features, and features were considered redundant if they shared a correlation coefficient of >0.9. Among the redundant features, the feature with the greatest variance was selected for unsupervised clustering.

Risk score development using machine learning

For an unbiased evaluation, the internal exploratory dataset was split into training and test datasets at a 7:3 ratio. The LASSO algorithm was adopted to select the most relevant radiomic features for predicting the proposed radiomic cluster. Standard 10-fold cross-validation was used in the regression to tune the parameters for the risk score construction. After model construction, the score was formulated as follows

$$S = \beta_0 + \beta 1 x_1 + \beta_2 x_2 + \ \ldots \ + \beta_k x_k$$

β_0 is the intercept from the LASSO model, β_{1-k} are the nonzero coefficients, and x_{1-k} are the feature values for the observation. The final performance was evaluated in the test dataset using the AUC metrics.

RNA sequencing and transcriptome analysis

RNA was isolated from the tissue samples using TRIzol (Invitrogen, Waltham, MA, USA). The concentration of the extracted RNA was determined using the Quant-IT RiboGreen assay (no. R11490, Thermo Fisher Scientific, Waltham, MA, USA), and RNA quality was evaluated using a TapeStation RNA ScreenTape (no. 5067-5576, Agilent Technologies, Santa Clara, CA, USA). Only RNA with an RNA integrity number of ≥7.0 was selected for subsequent library preparation. For each sample, 1 μg of RNA was used to prepare libraries using the Illumina TruSeq Stranded mRNA Sample Prep Kit (no. RS-122-2101, Illumina, Inc., San Diego, CA, USA), which included the initial step of mRNA isolation using poly(T)-attached magnetic beads. The mRNA was subsequently fragmented using divalent cations at high temperatures. The fragmented mRNA was converted into first-strand cDNA using SuperScript II reverse transcriptase (no. 18064014, Thermo Fisher Scientific) and random primers, followed by synthesis of second-strand cDNA using DNA polymerase I, ribonuclease H, and deoxyuridine triphosphate. The cDNA was then processed for end repair, A-tailing, adapter ligation, and enrichment by PCR. The final cDNA library was quantified using KAPA Library Quantification Kits (no. KK4854, Kapa Biosystems, Wilmington, MA, USA) and assessed for quality using TapeStation D1000 ScreenTape (no. 5067-5582, Agilent Technologies). Indexed libraries were sequenced on an Illumina NovaSeq 6000 platform to generate paired-end reads [2 × 100 bp (base pairs)]. Postsequencing data quality was verified using FastQC v0.11.7, and reads were cleaned of adapters and low-quality sequences using Trimmomatic 0.38. The cleaned reads were aligned to the reference genome GRCh37 (hg19) using HISAT2 v2.1.0, and transcripts were reconstructed using StringTie v2.1.3b. The read counts and fragment per kilobase of transcript per million mapped reads (FPKM) were obtained. The list of primers used for qPCR for the validation samples is provided in data S6.

GenePattern, an open web server for bioinformatics, was used for the downstream analysis of bulk transcriptomic data (*59*). DEG analysis and GSEA were performed with default parameters, as outlined in the software documentation. DEG profiles were configured based on gene set lists from the molecular signature database (*29*, *60*).

IHC staining

Formalin-fixed, paraffin-embedded tissue blocks were sectioned at 4-μm thickness using a microtome to generate nonstained slides. The slides were deparaffinized in xylene and rehydrated using a graded series of ethanol solutions. Antigens were retrieved using the heat-induced epitope retrieval method. The slides were immersed in tris-EDTA buffer (pH 9.0) and heated in a steamer for 20 min at 95°C. Following heat treatment, the slides were allowed to cool to room temperature for 30 min to facilitate the proper unfolding of the epitopes. The slides were then incubated with primary antibodies against ATF5 and PAX5. The ATF5 antibody (no. ab184923, Abcam, Cambridge, UK) was diluted 1:1000, and the PAX5 antibody (no. ab109443, Abcam) was diluted 1:100. After incubation with the primary antibody, the slides were washed three times with phosphate-buffered saline (PBS) to remove unbound antibodies. Subsequently, a secondary antibody, goat anti-rabbit immunoglobulin G H&L (no. ab205718, Abcam), was applied at 1:500 dilution. The slides were then incubated with secondary antibodies for 1 hour at room temperature in a humidified chamber,

followed by three additional washes with PBS to eliminate excess secondary antibodies.

For image analysis, four images per slide were captured from different regions using the LoupLite program at 40× magnification. Considering the use of the two antibodies, images were consistently obtained from the same regions for each analysis to ensure normalization across slides. The border areas of the slides were excluded to minimize noise from the background staining. Each image consisted of 6,291,456 pixels (3072 × 2048 resolution). Images were processed using ImageJ software, and the IHC Profiler plugin was used to analyze the pixel intensity (*61*). The percentage of positive staining was quantified by calculating the proportion of pixels corresponding to different intensity levels, thereby providing a robust assessment of protein expression.

Identification of BRAF, RAS, and TERT promoter mutations

DNA extraction was performed using the QIAamp DNA Mini Kit (QIAGEN, Inc., Hilden, Germany) according to the manufacturer's instructions. Genomic DNA was then amplified by PCR using the primers detailed in data S6 on a C1000 Thermal Cycler (Bio-Rad Laboratories, Hercules, CA, USA). After electrophoresis on a 2% agarose gel, the products were visualized using the Gel Doc EZ System (Bio-Rad) and purified using the QIAquick Gel Extraction Kit (QIAGEN, Inc., Hilden, Germany). Sequencing was conducted on an ABI 3730XL DNA Analyzer using the BigDye Terminator v3.1 Cycle Sequencing Ready Reaction Kit (Applied Biosystems, Waltham, MA, USA). deFuse v0.8.1, FusionCatcher v1.00, and Arriba v1.2.0 were applied to expression data to detect fusion oncogenes. Only the results consistently identified by all three analyses were considered significant (*62*–*64*).

Statistical analysis

All statistical analyses were performed using the R program v4.3.2 (R project, The R Foundation for Statistical Computing, Vienna, Austria) or GraphPad Prism v10 (GraphPad Software, San Diego, CA, USA). Continuous variables were compared across multiple groups using analysis of variance (ANOVA) and between two groups using an unpaired *t* test. The proportion of categorical variables was evaluated for independence using the chi-square test. Heatmaps and dendrograms were plotted using the "ComplexHeatmap" R package. The potential multicollinearity of the variables used for ordinal regression analysis was confirmed by calculating the variance inflation factor (VIF). VIF >5 was considered to indicate significant multicollinearity, and the results are provided in data S7. To ensure the robustness of our regression model, we systematically evaluated and addressed multicollinearity among all clinical biomarkers included in the analysis. The AUC of models developed in our regression analysis were compared using the DeLong test of the "pROC" R package. Gene signature–based survival analysis in patients with PTC from TCGA was conducted using GEPIA2 (http://gepia2.cancer-pku.cn) (*65*). Statistical significance was set at $P < 0.05$.

Study approval

The study protocol was approved by the Institutional Review Board (IRB) of Yonsei Cancer Center, Severance Hospital (IRB nos. 4-2021-1487 and 4-2013-0546), Seoul, South Korea. Written informed consent was obtained from all participants.

Supplementary Materials

The PDF file includes:
Figs. S1 to S5
Legends for data S1 to S8

Other Supplementary Material for this manuscript includes the following:
Data S1 to S8

REFERENCES AND NOTES

1. H. S. Ahn, H. J. Kim, K. H. Kim, Y. S. Lee, S. J. Han, Y. Kim, M. J. Ko, J. P. Brito, Thyroid cancer screening in South Korea increases detection of papillary cancers with no impact on other subtypes or thyroid cancer mortality. *Thyroid* **26**, 1535–1540 (2016).
2. J. Krajewska, A. Kukulska, M. Oczko-Wojciechowska, A. Kotecka-Blicharz, K. Drosik-Rutowicz, M. Haras-Gil, B. Jarzab, D. Handkiewicz-Junak, Early diagnosis of low-risk papillary thyroid cancer results rather in overtreatment than a better survival. *Front. Endocrinol. (Lausanne)* **11**, 571421 (2020).
3. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
4. B. Y. Cho, H. S. Choi, Y. J. Park, J. A. Lim, H. Y. Ahn, E. K. Lee, K. W. Kim, K. H. Yi, J. K. Chung, Y. K. Youn, N. H. Cho, D. J. Park, C. S. Koh, Changes in the clinicopathological characteristics and outcomes of thyroid cancer in Korea over the past four decades. *Thyroid* **23**, 797–804 (2013).
5. M. E. Cabanillas, D. G. McFadden, C. Durante, Thyroid cancer. *Lancet* **388**, 2783–2795 (2016).
6. S. Vaccarella, S. Franceschi, F. Bray, C. P. Wild, M. Plummer, L. Dal Maso, Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. *N. Engl. J. Med.* **375**, 614–617 (2016).
7. R. M. Tuttle, J. Fagin, G. Minkowitz, R. Wong, B. Roman, S. Patel, B. Untch, I. Ganly, A. Shaha, J. Shah, D. Li, A. Bach, J. Girshman, O. Lin, M. Cohen, J. M. Cohen, J. Cracchiolo, R. Ghossein, M. Sabra, L. Boucai, S. Fish, L. Morris, Active surveillance of papillary thyroid cancer: Frequency and time course of the six most common tumor volume kinetic patterns. *Thyroid* **32**, 1337–1345 (2022).
8. B. Altshuler, A. Bikas, T. Pappa, E. Marqusee, N. L. Cho, M. A. Nehs, J. B. Liu, G. M. Doherty, I. Landa, S. Ahmadi, E. K. Alexander, Nonoperative, active surveillance of larger malignant and suspicious thyroid nodules. *J. Clin. Endocrinol. Metab.* **109**, 1996–2002 (2024).
9. A. S. Ho, S. Kim, C. Zalt, M. L. Melany, I. E. Chen, J. Vasquez, J. Mallen-St Clair, M. M. Chen, M. Vasquez, X. Fan, W. K. van Deen, R. W. Haile, T. J. Daskivich, Z. S. Zumsteg, G. D. Braunstein, W. L. Sacks, Expanded parameters in active surveillance for low-risk papillary thyroid carcinoma: A nonrandomized controlled trial. *JAMA Oncol.* **8**, 1588–1596 (2022).
10. H. G. Welch, G. M. Doherty, Saving thyroids - Overtreatment of small papillary cancers. *N. Engl. J. Med.* **379**, 310–312 (2018).
11. I. Landa, T. Ibrahimpasic, L. Boucai, R. Sinha, J. A. Knauf, R. H. Shah, S. Dogan, J. C. Ricarte-Filho, G. P. Krishnamoorthy, B. Xu, N. Schultz, M. F. Berger, C. Sander, B. S. Taylor, R. Ghossein, I. Ganly, J. A. Fagin, Genomic and transcriptomic hallmarks of poorly differentiated and anaplastic thyroid cancers. *J. Clin. Invest.* **126**, 1052–1066 (2016).
12. Cancer Genome Atlas Research Network, Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676–690 (2014).
13. W. K. L. Doolittle, S. Park, S. G. Lee, S. Jeong, G. Lee, D. Ryu, K. Schoonjans, J. Auwerx, J. Lee, Y. S. Jo, Non-genomic activation of the AKT-mTOR pathway by the mitochondrial stress response in thyroid cancer. *Oncogene* **41**, 4893–4904 (2022).
14. M. Xing, A. S. Alzahrani, K. A. Carson, D. Viola, R. Elisei, B. Bendlova, L. Yip, C. Mian, F. Vianello, R. M. Tuttle, E. Robenshtok, J. A. Fagin, E. Puxeddu, L. Fugazzola, A. Czarniecka, B. Jarzab, C. J. O'Neill, M. S. Sywak, A. K. Lam, G. Riesco-Eizaguirre, P. Santisteban, H. Nakayama, R. P. Tufano, S. I. Pai, M. A. Zeiger, W. H. Westra, D. P. Clark, R. Clifton-Bligh, D. Sidransky, P. W. Ladenson, V. Sykorova, Association between BRAF V600E mutation and mortality in patients with papillary thyroid cancer. *JAMA* **309**, 1493–1501 (2013).
15. X. Yang, J. Li, X. Li, Z. Liang, W. Gao, J. Liang, S. Cheng, Y. Lin, TERT promoter mutation predicts radioiodine-refractory character in distant metastatic differentiated thyroid cancer. *J. Nucl. Med.* **58**, 258–265 (2017).
16. P. Yu, N. Qu, R. Zhu, J. Hu, P. Han, J. Wu, L. Tan, H. Gan, C. He, C. Fang, Y. Lei, J. Li, C. He, F. Lan, X. Shi, W. Wei, Y. Wang, Q. Ji, F. X. Yu, Y. L. Wang, TERT accelerates BRAF mutant-induced thyroid cancer dedifferentiation and progression by regulating ribosome biogenesis. *Sci. Adv.* **9**, eadg7125 (2023).
17. T. Carling, R. Udelsman, Thyroid cancer. *Annu. Rev. Med.* **65**, 125–137 (2014).
18. Q. Zhao, J. Ming, C. Liu, L. Shi, X. Xu, X. Nie, T. Huang, Multifocality and total tumor diameter predict central neck lymph node metastases in papillary thyroid microcarcinoma. *Ann. Surg. Oncol.* **20**, 746–752 (2013).

19. W. K. Lee, J. Lee, H. Kim, S. G. Lee, S. H. Choi, S. Jeong, H. J. Kwon, S. G. Jung, Y. S. Jo, Peripheral location and infiltrative margin predict invasive features of papillary thyroid microcarcinoma. *Eur. J. Endocrinol.* **181**, 139–149 (2019).
20. Y. Liu, Y. Wang, K. Zhao, D. Li, Z. Chen, R. Jiang, X. Wang, X. He, Lymph node metastasis in young and middle-aged papillary thyroid carcinoma patients: A SEER-based cohort study. *BMC Cancer* **20**, 181 (2020).
21. L. J. DeGroot, E. L. Kaplan, M. McCormick, F. H. Straus, Natural history, treatment, and course of papillary thyroid carcinoma. *J. Clin. Endocrinol. Metab.* **71**, 414–424 (1990).
22. W. J. Wei, Z. W. Lu, D. Wen, T. Liao, D. S. Li, Y. Wang, Y. X. Zhu, Z. Y. Wang, Y. Wu, Y. L. Wang, Q. H. Ji, The positive lymph node number and postoperative N-staging used to estimate survival in patients with differentiated thyroid cancer: Results from the surveillance, epidemiology, and end results dataset (1988-2008). *World J. Surg.* **42**, 1762–1771 (2018).
23. D. H. Seo, S. G. Lee, H. Y. Lee, S. Jeong, S. Park, J. Lee, Y. S. Jo, Lymph node metastasis-dependent molecular classification in papillary thyroid carcinoma defines aggressive metastatic outgrowth. *Clin. Transl. Med.* **13**, e1211 (2023).
24. W. Pu, X. Shi, P. Yu, M. Zhang, Z. Liu, L. Tan, P. Han, Y. Wang, D. Ji, H. Gan, W. Wei, Z. Lu, N. Qu, J. Hu, X. Hu, Z. Luo, H. Li, Q. Ji, J. Wang, X. Zhang, Y. L. Wang, Single-cell transcriptomic analysis of the tumor ecosystems underlying initiation and progression of papillary thyroid carcinoma. *Nat. Commun.* **12**, 6058 (2021).
25. J. Yu, Y. Deng, T. Liu, J. Zhou, X. Jia, T. Xiao, S. Zhou, J. Li, Y. Guo, Y. Wang, J. Zhou, C. Chang, Lymph node metastasis prediction of papillary thyroid carcinoma based on transfer learning radiomics. *Nat. Commun.* **11**, 4807 (2020).
26. W. Lu, L. Zhong, D. Dong, M. Fang, Q. Dai, S. Leng, L. Zhang, W. Sun, J. Tian, J. Zheng, Y. Jin, Radiomic analysis for preoperative prediction of cervical lymph node metastasis in patients with papillary thyroid carcinoma. *Eur. J. Radiol.* **118**, 231–238 (2019).
27. H. Zhang, S. Hu, X. Wang, J. He, W. Liu, C. Yu, Z. Sun, Y. Ge, S. Duan, Prediction of cervical lymph node metastasis using MRI radiomics approach in papillary thyroid carcinoma: A feasibility study. *Technol. Cancer Res. Treat.* **19**, 1533033820969451 (2020).
28. I. Sugitani, Y. Ito, A. Miyauchi, T. Imai, S. Suzuki, Active surveillance versus immediate surgery: Questionnaire survey on the current treatment strategy for adult patients with low-risk papillary thyroid microcarcinoma in Japan. *Thyroid* **29**, 1563–1571 (2019).
29. A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, P. Tamayo, The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
30. C. Cobaleda, A. Schebesta, A. Delogu, M. Busslinger, Pax5: The guardian of B cell identity and function. *Nat. Immunol.* **8**, 463–470 (2007).
31. C. J. Fiorese, A. M. Schulz, Y. F. Lin, N. Rosin, M. W. Pellegrino, C. M. Haynes, The transcription factor ATF5 mediates a mammalian mitochondrial UPR. *Curr. Biol.* **26**, 2037–2043 (2016).
32. X. Gao, X. Ran, W. Ding, The progress of radiomics in thyroid nodules. *Front. Oncol.* **13**, 1109319 (2023).
33. J. Tang, S. Jiang, J. Ma, X. Xi, H. Li, L. Wang, B. Zhang, Nomogram based on radiomics analysis of ultrasound images can improve preoperative BRAF mutation diagnosis for papillary thyroid microcarcinoma. *Front. Endocrinol. (Lausanne)* **13**, 915135 (2022).
34. M. R. Kwon, J. H. Shin, H. Park, H. Cho, S. Y. Hahn, K. W. Park, Radiomics study of thyroid ultrasound for predicting BRAF Mutation in papillary thyroid carcinoma: preliminary results. *AJNR Am. J. Neuroradiol.* **41**, 700–705 (2020).
35. M. Xing, R. Liu, X. Liu, A. K. Murugan, G. Zhu, M. A. Zeiger, S. Pai, J. Bishop, BRAF V600E and TERT promoter mutations cooperatively identify the most aggressive papillary thyroid cancer with highest recurrence. *J. Clin. Oncol.* **32**, 2718–2726 (2014).
36. D. H. Seo, S. G. Lee, S. M. Choi, H. Y. Kim, S. Park, S. G. Jung, Y. S. Jo, J. Lee, Promoter mutation-independent TERT expression is related to immune-enriched milieu in papillary thyroid cancer. *Endocr. Relat. Cancer* **31**, e240068 (2024).
37. W. Li, Y. Wang, J. Wen, L. Zhang, Y. Sun, Diagnostic performance of american college of radiology TI-RADS: A systematic review and meta-analysis. *AJR Am. J. Roentgenol.* **216**, 38–47 (2021).
38. D. H. Kim, S. R. Chung, S. H. Choi, K. W. Kim, Accuracy of thyroid imaging reporting and data system category 4 or 5 for diagnosing malignancy: A systematic review and meta-analysis. *Eur. Radiol.* **30**, 5611–5624 (2020).
39. Q. Qi, A. Zhou, S. Guo, X. Huang, S. Chen, Y. Li, P. Xu, Explore the diagnostic efficiency of Chinese thyroid imaging reporting and data systems by comparing with the other four systems (ACR TI-RADS, Kwak-TIRADS, KSThR-TIRADS, and EU-TIRADS): A single-center study. *Front. Endocrinol. (Lausanne)* **12**, 763897 (2021).
40. S. H. Choi, E. K. Kim, J. Y. Kwak, M. J. Kim, E. J. Son, Interobserver and intraobserver variations in ultrasound assessment of thyroid nodules. *Thyroid* **20**, 167–172 (2010).
41. S. H. Kim, C. S. Park, S. L. Jung, B. J. Kang, J. Y. Kim, J. J. Choi, Y. I. Kim, J. K. Oh, J. S. Oh, H. Kim, S. H. Jeong, H. W. Yim, Observer variability and the performance between faculties and residents: US criteria for benign and malignant thyroid nodules. *Korean J. Radiol.* **11**, 149–155 (2010).
42. R. J. Gillies, P. E. Kinahan, H. Hricak, Radiomics: Images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).
43. Y. Ito, A. Miyauchi, Active surveillance of low-risk papillary thyroid microcarcinomas in Japan and other countries: A review. *Expert. Rev. Endocrinol. Metab.* **15**, 5–12 (2020).
44. I. Sugitani, Y. Ito, D. Takeuchi, H. Nakayama, C. Masaki, H. Shindo, M. Teshima, K. Horiguchi, Y. Yoshida, T. Kanai, M. Hirokawa, K. Y. Hames, I. Tabei, A. Miyauchi, Indications and strategy for active surveillance of adult low-risk papillary thyroid microcarcinoma: Consensus statements from the Japan association of endocrine surgery task force on management for papillary thyroid microcarcinoma. *Thyroid* **31**, 183–192 (2021).
45. G. Orlando, G. Scerrino, A. Corigliano, I. Vitale, R. Tutino, S. Radellini, F. Cupido, G. Graceffa, G. Cocorullo, G. Salamone, G. Melfa, Papillary thyroid microcarcinoma: Active surveillance against surgery. Considerations of an Italian working group from a systematic review. *Front. Oncologia* **12**, 859461 (2022).
46. A. I. F. Poon, J. J. Y. Sung, Opening the black box of AI-medicine. *J. Gastroenterol. Hepatol.* **36**, 581–584 (2021).
47. M. R. Karim, T. Islam, M. Shajalal, O. Beyan, C. Lange, M. Cochez, D. Rebholz-Schuhmann, S. Decker, Explainable AI for bioinformatics: Methods, tools and applications. *Brief Bioinform.* **24**, bbad236 (2023).
48. J. Lee, J. H. Yoon, E. Lee, H. Y. Lee, S. Jeong, S. Park, Y. S. Jo, J. Y. Kwak, Immune response and mesenchymal transition of papillary thyroid carcinoma reflected in ultrasonography features assessed by radiologists and deep learning. *J. Adv. Res.* **62**, 219–228 (2024).
49. S. Leboulleux, R. M. Tuttle, F. Pacini, M. Schlumberger, Papillary thyroid microcarcinoma: Time to shift from surgery to active surveillance? *Lancet Diabetes Endocrinol.* **4**, 933–942 (2016).
50. A. de Biase, N. Sourlos, P. M. A. van Ooijen, Standardization of artificial intelligence development in radiotherapy. *Semin. Radiat. Oncol.* **32**, 415–420 (2022).
51. S. Y. Kim, J. Y. Kwak, E. K. Kim, J. H. Yoon, H. J. Moon, Association of preoperative US features and recurrence in patients with classic papillary thyroid carcinoma. *Radiology* **277**, 574–583 (2015).
52. S. Y. Nam, J. H. Shin, B. K. Han, E. Y. Ko, E. S. Ko, S. Y. Hahn, J. H. Chung, Preoperative ultrasonographic features of papillary thyroid carcinoma predict biological behavior. *J. Clin. Endocrinol. Metab.* **98**, 1476–1482 (2013).
53. C. Durante, L. Hegedüs, D. G. Na, E. Papini, J. A. Sipos, J. H. Baek, A. Frasoldati, G. Grani, E. Grant, E. Horvath, J. K. Hoang, S. J. Mandel, W. D. Middleton, R. Ngu, L. A. Orloff, J. H. Shin, P. Trimboli, J. H. Yoon, F. N. Tessler, International expert consensus on US lexicon for thyroid nodules. *Radiology* **309**, e231481 (2023).
54. B. R. Haugen, E. K. Alexander, K. C. Bible, G. M. Doherty, S. J. Mandel, Y. E. Nikiforov, F. Pacini, G. W. Randolph, A. M. Sawka, M. Schlumberger, K. G. Schuff, S. I. Sherman, J. A. Sosa, D. L. Steward, R. M. Tuttle, L. Wartofsky, 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: The American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid* **26**, 1–133 (2016).
55. E. K. Kim, C. S. Park, W. Y. Chung, K. K. Oh, D. I. Kim, J. T. Lee, H. S. Yoo, New sonographic criteria for recommending fine-needle aspiration biopsy of nonpalpable solid nodules of the thyroid. *AJR Am. J. Roentgenol.* **178**, 687–691 (2002).
56. J. Y. Kwak, K. H. Han, J. H. Yoon, H. J. Moon, E. J. Son, S. H. Park, H. K. Jung, J. S. Choi, B. M. Kim, E. K. Kim, Thyroid imaging reporting and data system for US features of nodules: A step in establishing better stratification of cancer risk. *Radiology* **260**, 892–899 (2011).
57. D. W. Kim, C. K. Eun, H. S. In, M. H. Kim, S. J. Jung, S. K. Bae, Sonographic differentiation of asymptomatic diffuse thyroid disease from normal thyroid: A prospective study. *AJNR Am. J. Neuroradiol.* **31**, 1956–1960 (2010).
58. F. N. Tessler, W. D. Middleton, E. G. Grant, Thyroid imaging reporting and data system (TI-RADS): A user's guide. *Radiology* **287**, 29–36 (2018).
59. M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, J. P. Mesirov, GenePattern 2.0. *Nat. Genet.* **38**, 500–501 (2006).
60. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).
61. F. Varghese, A. B. Bukhari, R. Malhotra, A. De, IHC Profiler: An open source plugin for the quantitative evaluation and automated scoring of immunohistochemistry images of human tissue samples. *PLOS ONE* **9**, e96801 (2014).
62. A. McPherson, F. Hormozdiari, A. Zayed, R. Giuliany, G. Ha, M. G. Sun, M. Griffith, A. Heravi Moussavi, J. Senz, N. Melnyk, M. Pacheco, M. A. Marra, M. Hirst, T. O. Nielsen, S. C. Sahinalp, D. Huntsman, S. P. Shah, deFuse: An algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.* **7**, e1001138 (2011).
63. D. Nicorici, M. Şatalan, H. Edgren, S. Kangaspeska, A. Murumägi, O. Kallioniemi, S. Virtanen, O. Kilkku, FusionCatcher – A tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*, 011650 (2014).

64. S. Uhrig, J. Ellermann, T. Walther, P. Burkhardt, M. Fröhlich, B. Hutter, U. H. Toprak, O. Neumann, A. Stenzinger, C. Scholl, S. Fröhling, B. Brors, Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **31**, 448–460 (2021).
65. Z. Tang, B. Kang, C. Li, T. Chen, Z. Zhang, GEPIA2: An enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* **47**, W556–W560 (2019).

Acknowledgments: We thank J. Y. Kim and H. C. Yu for technical support. We also thank Medical Illustration and Design, part of the Medical Research Support Services of Yonsei University College of Medicine, for all the artistic support related to this work. **Funding:** We received funding from the National Research Foundation of Korea (NRF), which is funded by the Korean government: NRF-2018R1C1B6006064 (J.H.Y.), NRF-2020R1A2C1006047 (J.L.), NRF-2021R1A2C2007492 (J.Y.K.), and NRF-2023R1A2C1003167 (Y.S.J.). **Author contributions:** Conceptualization: D.H.S., J.L., J.Y.K., and Y.S.J. Methodology: D.H.S., E.L., J.H.Y., E.G.P., S.P., H.Y.L., and K.H. Software: D.H.S., E.L., and E.G.P. Validation: D.H.S., E.L., J.H.Y., S.P., and K.H. Formal analysis: D.H.S., E.G.P., and K.H. Investigation: D.H.S. and J.H.Y. Writing—original draft: D.H.S., J.H.Y., and E.G.P. Writing—review and editing: E.L., J.H.Y., S.P., J.L., J.Y.K., and Y.S.J. Visualization: D.H.S., E.L., and E.G.P. Supervision: S.P., H.Y.L., J.H., C.R.L., J.L., J.Y.K., and Y.S.J. Project administration: J.H.Y., J.H., C.R.L., J.L., J.Y.K., and Y.S.J. Funding acquisition: J.L., J.Y.K., and Y.S.J. All authors have read and approved the final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The key clinical and corresponding transcriptomic datasets generated and/or analyzed during this study are available at the EMBL-EBI under accession number E-MTAB-15233 (www.ebi.ac.uk/biostudies/ArrayExpress/studies/E-MTAB-15233). All analyses were performed using publicly available R packages, as detailed in the methodology section, to ensure the reproducibility of the computational workflow. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 3 January 2025
Accepted 29 July 2025
Published 29 August 2025
10.1126/sciadv.adv6697