Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Improving specificity in high-sensitivity NSCLC PET/CT nodal classification with patient-specific metabolic context

Resit Akyel
Department of Nuclear Medicine, Yedikule Chest Diseases and Thoracic Surgery Training and Research Hospital

Muhammed Mustafa Atci
Department of Medical Oncology, Istinye University Liv Hospital Topkapi

Fatma Arzu Gortan
Ankara Bilkent City Hospital

Demet Turan
Department of Chest Diseases, Yedikule Chest Diseases and Thoracic Surgery Training and Research Hospital

Isa Yildirim
iyildirim@itu.edu.tr

Department of Electronics and Communication Engineering, Istanbul Technical University

Research Article

Keywords: non-small cell lung cancer, FDG-PET/CT, mediastinal lymph nodes, false-positive classification, SUVmax, metabolic context, prediction models

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11655820/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

High-sensitivity [18F]FDG-PET/CT assessment of thoracic lymph nodes in non-small cell lung cancer (NSCLC) can generate false-positive classifications when benign inflammatory uptake resembles metastatic disease. We investigated whether interpreting each nodal station within the patient's wider intrathoracic metabolic pattern can refine decisions based on local SUVmax alone.

Methods

We retrospectively studied 486 patients and 1,111 pathology-confirmed outcomes at six hilar or mediastinal nodal stations. Each station was assessed using local SUVmax-only logistic regression or contextual models incorporating uptake in other nodal stations, mediastinal blood-pool activity, primary-tumor uptake and laterality, and recorded inflammatory variables. Elastic-net logistic regression (EN), histogram gradient boosting, and radial-basis support-vector machines underwent ten repetitions of patient-grouped nested cross-validation. Thresholds targeting 90% or 95% sensitivity were selected from training data and transferred unchanged to held-out patients. An EN-versus-SUVmax global operating-point analysis was performed post hoc.

Results

At the 95% training-derived sensitivity target, an exploratory pooled comparison showed higher specificity with contextual elastic-net modeling than with SUVmax-only logistic regression (82.48% vs 78.20%; difference + 4.28 percentage points; conditional 95% confidence interval [CI] + 1.86 to + 6.76). Positive predictive value increased from 75.66% to 79.31% (+ 3.65 points; CI + 1.47 to + 5.97), while achieved sensitivity was 94.20% versus 95.13% (-0.93 points; CI -2.36 to + 0.45). Equal-station averaging supported the same pattern. The specificity gain was concentrated at the more stringent operating target.

Conclusions

Integrating intrathoracic metabolic measurements already available on routine PET/CT yielded more specific nodal classification under a stringent high-sensitivity operating rule. These results identify a practical opportunity for context-aware station assessment alongside local SUVmax and provide a clear basis for independent clinical validation.

Background

Accurate intrathoracic nodal assessment is central to staging and treatment planning in non-small cell lung cancer (NSCLC). [18F]FDG-PET/CT contributes essential metabolic and anatomical information, but FDG avidity also arises from infection, granulomatous disease, and other inflammatory processes. Distinguishing these causes of uptake is particularly important when PET/CT findings may influence staging pathways and require tissue confirmation [1–7].

The challenge is most apparent when a high-sensitivity interpretation is required. Lower classification thresholds capture more potentially malignant stations but also label more benign stations as suspicious. A method that refines specificity at a clinically relevant high-sensitivity operating point could add value to PET/CT interpretation. Importantly, much of the information needed for a richer assessment is already present in the same examination.

Previous NSCLC investigations have examined nodal SUV measurements, machine learning, radiomics, multimodality fusion, station-specific prediction, and automated nodal detection [8–18]. These studies demonstrate the potential of quantitative PET/CT, but they also make a strong baseline comparator essential: local SUVmax remains readily available, interpretable, and highly discriminative. The less explored question is whether metabolic findings elsewhere in the thorax contribute useful station-level information beyond that local measurement.

We assessed this question using patient-specific intrathoracic metabolic context, including uptake in other hilar and mediastinal stations, mediastinal blood-pool activity, primary-tumor SUVmax and laterality, and recorded inflammatory findings. Three contextual approaches were compared with station-specific SUVmax-only logistic regression using repeated, patient-grouped nested validation. We then characterized the clinical operating profile of contextual elastic-net modeling at training-derived high-sensitivity thresholds, including an exploratory, post hoc pooled analysis across stations.

Methods

Study design, cohort assembly, and reference standard

We retrospectively analyzed a single-center clinical PET/CT dataset with linked nodal pathology. The source contained 506 records with pathology entries for seven nodal stations. Records coded as having bilateral primary tumors were excluded, and histologic eligibility was restricted to the recorded NSCLC categories (codes 1–3). This yielded 488 records associated with 487 patient identifiers. Two nonidentical records had the same identifier but discordant station-7 pathology. Because neither could be adjudicated from the available analytical source, both were removed before modeling; the final cohort comprised 486 unique patients.

The observation unit for prediction was a patient–station pair. Malignant and benign status were defined only from station-specific pathology (codes 1 and 0, respectively). Unverified stations (code 4) and other unknown outcomes were excluded for the corresponding station rather than assigned a benign label. Clinical follow-up was not used to create surrogate pathology outcomes. The final modeled cohort

contained 1,111 tissue-confirmed binary outcomes from six stations; station 2L was not modeled because its 15 verified results were all benign (Fig. 1).

Pathological verification was completed within 30 days after PET/CT. If surgical and endobronchial ultrasound-guided transbronchial needle aspiration (EBUS-TBNA) pathology were both available for the same station, surgical pathology defined the reference label. Where surgical sampling was absent, an adequately sampled EBUS-TBNA result defined the label. Adequate negative EBUS-TBNA samples were classified as non-metastatic; nondiagnostic or inadequate samples were excluded. These rules were applied at the station level. Because not all stations underwent tissue verification, the modeled population represents verified stations and remains vulnerable to selection and partial verification bias.

PET/CT protocol and metabolic measurements

PET/CT protocol and metabolic measurements
Patients fasted for at least 6 hours before intravenous [18F]FDG injection (3.7 MBq/kg), with blood glucose below 150 mg/dL at administration. Imaging followed an uptake interval of approximately 60 minutes on a Discovery 600 PET/CT scanner (GE Medical Systems, Milwaukee, WI, USA). Noncontrast low-dose CT (120 kV, 100–210 mAs) supplied attenuation correction and anatomical localization. Three-dimensional PET extended from the skull base to the mid-thigh at 1.75 minutes per bed position; standardized uptake values (SUVs) were body-weight normalized [19].

Maximum SUV (SUVmax) was measured at stations 10R, 10L, 7, 4R, 4L, 2R, and 2L using established anatomical definitions [1]. When several nodes were visible within a station, the volume of interest was placed over the most avid node while avoiding neighboring tumor or structures. When no discrete avid node was identified, the volume of interest was placed within the anatomical station. Blood-pool SUVmax was measured in the ascending-aortic lumen with care to avoid the wall and adjacent tissues. Two experienced nuclear medicine physicians jointly reviewed images and reached consensus on station identity, volume placement, and SUVmax. Because readings were not independent, no interobserver-reliability estimate was available.

Predictors, comparator, and contextual models

The 12 recorded contextual candidate variables were primary-tumor SUVmax; mediastinal aortic blood-pool SUVmax; an indicator of pneumonia/consolidation and its SUVmax; SUVmax at stations 10R, 10L, 7, 4R, 4L, 2R, and 2L; and primary-tumor laterality (Table 1). The station-specific target SUVmax was therefore among the contextual inputs. Histology informed cohort selection but was not used as a model predictor. The pneumonia/consolidation SUVmax variable was missing in 73.7% of records; this motivated a separate ablation analysis.

Table 1
Intrathoracic-context predictors and missingness. Missingness is reported for the 486 primary-analysis records. Histologic category was used for eligibility but not as a predictor.

Column	Predictor	Missing n (%)
C	Primary-tumor SUVmax	2 (0.4%)
D	Mediastinal aortic blood-pool SUVmax	0 (0.0%)
E	Pneumonia/consolidation indicator	13 (2.7%)
F	SUVmax for pneumonia/consolidation	358 (73.7%)
G-L	Station SUVmax: 10R, 10L, 7, 4R, 4L, 2R	0 (0.0%)
M	Station 2L SUVmax	1 (0.2%)
U	Primary-tumor side (right 0, left 1)	0 (0.0%)

The benchmark was logistic regression based solely on the SUVmax at the target station (SUV_LR). Three contextual methods were evaluated: elastic-net logistic regression (EN), histogram gradient boosting (HGB), and a radial-basis-function support-vector machine (SVM). Within training partitions, EN and SVM used median imputation, missing-value indicators where applicable, and standardization; HGB used native missing-value handling. The strict local-SUVmax comparator allowed estimation of incremental predictive information rather than an undifferentiated machine-learning comparison.

Patient-grouped nested validation and clinical operating thresholds

Each nodal station was modeled separately. Five-fold outer cross-validation was combined with three-fold inner, patient-grouped validation and repeated for ten fixed random seeds (3101–3110). Matching outer partitions were used across models for the same station and repetition. Inner-fold area under the receiver operating characteristic curve (ROC-AUC) determined hyperparameters. Imputation, scaling, fitting, probability calibration, and threshold selection were confined to the relevant training data [20]. Neither oversampling, class weighting, synthetic data generation, nor outcome imputation was performed.

Training-derived calibrated thresholds targeted at least 90% or 95% sensitivity: within each outer training fold, the highest threshold satisfying the target on inner out-of-fold training predictions was selected. The threshold was then applied unchanged to held-out outer-fold predictions. We averaged repetition-level sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and predicted-negative fraction. Because each model had its own training-derived cutoff, achieved held-out sensitivities need not be exactly equal. ROC-AUC and Brier score were also reported independent of the high-sensitivity cutoffs.

After reviewing heterogeneous station-level findings, we performed an exploratory global EN-versus-SUV_LR comparison using the retained, held-out predictions. For each repetition, micro-averaging pooled verified observations from all six modeled stations; equal-station macro-averaging assigned each station equal weight. Paired patient-cluster bootstrap intervals (5,000 draws) preserved within-patient correlation across stations. These intervals condition on the retained predictions and do not account for full model-refitting uncertainty. Because the global EN contrast was selected post hoc, its intervals are descriptive rather than confirmatory.

Analyses of robustness, verification, and feature reliance

Station-by-model differences were examined with paired patient-cluster bootstrap intervals. An explanatory matched-sensitivity analysis reselected thresholds on retained held-out outcomes to attain sensitivity at or above 95% as closely as possible. This procedure necessarily uses test outcomes, so its results assess the interpretation of the initial operating-point comparison; they do not constitute an implementable or independently validated decision rule.

ROC-AUC and Brier-score comparisons across three contextual models and six stations formed a 36-comparison family. Specificity and PPV comparisons at 90% and 95% targets formed a 72-comparison family. We applied Nadeau–Bengio dependence-corrected paired fold-level inference and Holm multiplicity adjustment [21, 22]. The corrected analyses are presented alongside, rather than replaced by, conditional bootstrap intervals.

Calibration was summarized using calibration-in-the-large, joint calibration intercept, and calibration slope [23]. Decision curves evaluated net benefit across illustrative probability thresholds, with no claim that a specific clinical threshold was prespecified [24]. To assess reliance on recorded inflammatory information, both pneumonia/consolidation variables were removed together while retaining matched partitions and full-context hyperparameters.

Potential partial verification bias was explored using patient-grouped cross-fitted estimates of pathological verification probability. Inverse-probability weights, with truncation of extreme values, were used for sensitivity analyses under a missing-at-random assumption, not as a definitive recovery of unverified pathological truth [25, 26]. Within held-out folds, grouped permutation importance jointly shuffled other-station uptake, blood-pool activity, primary-tumor uptake, laterality, or inflammation variables; repeated shuffles estimated changes in ROC-AUC. Reporting was informed by TRIPOD + AI and relevant guidance on the evaluation and reporting of medical machine-learning prediction models [27, 28].

Ethics

The study protocol was approved by the Non-Interventional Clinical Research Ethics Committee of Istanbul Yedikule Chest Diseases and Thoracic Surgery Training and Research Hospital (decision no.

2026-10, dated 26 February 2026). Informed consent was obtained from all participants. The study was conducted in accordance with the Declaration of Helsinki and applicable institutional guidelines and regulations.

Results

Pathology-confirmed cohort and the SUVmax benchmark

The 486 patients contributed 1,111 pathology-confirmed observations: 462 malignant and 649 benign. The six modeled stations comprised 10R (161 observations, 69 malignant), 10L (108, 42), station 7 (353, 147), 4R (283, 138), 4L (150, 45), and 2R (56, 21). Station 2L had only benign verified results and was not modeled. Sample size varied substantially across stations, as did the amount of pathology verification (Table 2).

Local SUVmax was a demanding comparator, with station-specific SUV_LR ROC-AUC values of approximately 0.939–0.967 (Table 2). Against this strong baseline, the central question was whether information distributed across the patient's thorax could sharpen high-sensitivity classification of individual nodal stations.

Table 2
Verified station samples and SUVmax benchmark. AUC intervals are conditional on retained predictions and are not multiplicity-adjusted.

Station	Verified	Malignant	Unknown	SUV_LR AUC (95% CI)	Brier
10R	161	69	325	0.9394 (0.8972, 0.9776)	0.0857
10L	108	42	378	0.9509 (0.9011, 0.9869)	0.0954
7	353	147	132	0.9661 (0.9466, 0.9829)	0.0700
4R	283	138	203	0.9442 (0.9113, 0.9728)	0.0864
4L	150	45	336	0.9577 (0.9170, 0.9883)	0.0691
2R	56	21	430	0.9667 (0.9114, 0.9984)	0.0816
2L	15	0	470	Not estimable	Not fitted

Higher specificity at a stringent high-sensitivity operating point

The clearest operating-point signal emerged at the training-derived 95% sensitivity target (Table 3; Fig. 2). In an exploratory post hoc pooled comparison across the six verified stations, contextual elastic-net logistic regression (EN) increased specificity from 78.20% with local SUVmax-only logistic regression (SUV_LR) to 82.48% (difference + 4.28 percentage points; conditional 95% CI + 1.86 to + 6.76). Positive predictive value (PPV) rose from 75.66% to 79.31% (+ 3.65 points; CI + 1.47 to + 5.97). Achieved

sensitivity was 94.20% versus 95.13% (-0.93 points; CI -2.36 to + 0.45), and negative predictive value (NPV) was 95.23% versus 95.76% (-0.52 points; CI -1.67 to + 0.61). Thus, the contextual approach yielded fewer false-positive classifications under the transferred high-sensitivity thresholds, with a small mean change in sensitivity.

Table 3
Global operating-point results at both sensitivity targets. Micro-averaging pools all pathology-verified station classifications within each repetition. Equal-station macro-averaging gives each modeled station equal weight. Differences are EN minus SUV_LR. Conditional 95% intervals use paired patient-cluster bootstrap resampling and do not include full model-refitting uncertainty.

Target	Weighting	Metric	SUV_LR %	EN %	Delta pp	95% CI for delta
90%	Micro	Sensitivity	89.96	89.74	-0.22	-1.86 to + 1.42
90%	Micro	Specificity	91.71	91.54	-0.17	-1.84 to + 1.57
90%	Micro	PPV	88.55	88.32	-0.23	-2.25 to + 1.90
90%	Micro	NPV	92.77	92.61	-0.16	-1.23 to + 0.94
90%	Macro	Sensitivity	90.12	89.46	-0.66	-2.47 to + 1.04
90%	Macro	Specificity	91.64	90.72	-0.92	-2.77 to + 0.85
90%	Macro	PPV	87.62	86.20	-1.43	-3.85 to + 0.77
90%	Macro	NPV	93.20	92.80	-0.40	-1.62 to + 0.73
95%	Micro	Sensitivity	95.13	94.20	-0.93	-2.36 to + 0.45
95%	Micro	Specificity	78.20	82.48	+ 4.28	+ 1.86 to + 6.76
95%	Micro	PPV	75.66	79.31	+ 3.65	+ 1.47 to + 5.97
95%	Micro	NPV	95.76	95.23	-0.52	-1.67 to + 0.61
95%	Macro	Sensitivity	95.16	93.84	-1.32	-3.31 to + 0.37
95%	Macro	Specificity	74.91	80.99	+ 6.09	+ 3.62 to + 8.55
95%	Macro	PPV	72.14	76.61	+ 4.46	+ 2.22 to + 6.84
95%	Macro	NPV	95.82	95.12	-0.70	-2.11 to + 0.54

The pattern was also present when all six stations received equal weight: macro-averaged specificity improved by 6.09 points (95% CI + 3.62 to + 8.55), and PPV by 4.46 points (+ 2.22 to + 6.84). Corresponding differences were − 1.32 points for sensitivity (-3.31 to + 0.37) and − 0.70 for NPV (-2.11 to + 0.54). The agreement between pooled and station-balanced summaries suggests that the 95%-target specificity shift was not solely a consequence of the largest stations. These descriptive intervals reflect retained predictions from an exploratory comparison.

Operating-target dependence and nodal station patterns

The operating strategy mattered. At the 90% training-sensitivity target, pooled EN–SUV_LR differences were small: -0.22 percentage points for sensitivity, -0.17 for specificity, and − 0.23 for PPV, with conditional intervals spanning zero (Table 3). The principal contextual gain was therefore concentrated in the more stringent 95% sensitivity setting, where avoiding additional false-positive station classifications is particularly relevant.

At the 95% target, the largest station-specific EN effects appeared at 10R and 4L (Tables 4 and 5). At 10R, specificity rose from 68.5% to 81.8% and PPV from 69.7% to 79.6%, while sensitivity changed from 95.4% to 94.2%. Paired differences were + 13.4 points for specificity (95% CI + 7.5 to + 19.8), + 10.0 for PPV (+ 5.1 to + 15.5), and − 1.2 for sensitivity (-4.8 to + 2.2). At 4L, specificity increased from 65.7% to 72.2% (+ 6.5; +1.5 to + 11.6) and PPV from 54.9% to 59.9% (+ 5.0; +1.0 to + 9.2), with similar achieved sensitivity. Station 7 had smaller favorable EN differences; other stations were mixed. The 2R findings were considered exploratory because only 56 verified outcomes were available.

Table 4

Station-specific results at the 95% sensitivity target. Values are means across ten repetitions. Each model uses its own training-selected threshold, so achieved test sensitivity can differ. Delta values are percentage-point changes relative to SUV_LR.

Station	Model	Sens. %	Spec. %	PPV %	Delta spec. pp	Delta PPV pp
10R	SUV_LR	95.4	68.5	69.7	-	-
10R	EN	94.2	81.8	79.6	+ 13.4	+ 10.0
10R	HGB	95.2	63.5	66.3	-5.0	-3.4
10R	SVM	94.3	77.3	76.0	+ 8.8	+ 6.3
10L	SUV_LR	94.5	85.0	80.4	-	-
10L	EN	94.0	85.6	80.7	+ 0.6	+ 0.3
10L	HGB	94.8	84.7	80.2	-0.3	-0.2
10L	SVM	94.3	80.5	75.5	-4.5	-4.8
7	SUV_LR	95.2	86.6	83.5	-	-
7	EN	94.5	89.8	86.9	+ 3.3	+ 3.4
7	HGB	95.4	88.3	85.3	+ 1.7	+ 1.8
7	SVM	95.2	84.3	81.2	-2.3	-2.2
4R	SUV_LR	95.1	82.6	83.9	-	-
4R	EN	94.2	79.9	81.8	-2.6	-2.1
4R	HGB	95.1	81.4	83.0	-1.2	-0.8
4R	SVM	94.2	73.1	77.0	-9.4	-6.9
4L	SUV_LR	95.1	65.7	54.9	-	-
4L	EN	94.7	72.2	59.9	+ 6.5	+ 5.0
4L	HGB	95.3	49.0	44.7	-16.7	-10.2
4L	SVM	95.1	66.7	55.2	+ 1.0	+ 0.2
2R	SUV_LR	95.7	61.1	60.6	-	-
2R	EN	91.4	76.6	70.8	+ 15.4	+ 10.2
2R	HGB	95.7	68.0	65.3	+ 6.9	+ 4.7
2R	SVM	93.3	54.9	55.8	-6.3	-4.8

Table 5
Paired station-specific operating-point differences. Differences are contextual model minus SUV_LR at the training-derived 95% sensitivity target. Intervals use paired patient-cluster bootstrap resampling, are conditional on retained predictions, and are not multiplicity-adjusted.

Station	Model	Delta sensitivity, pp (95% CI)	Delta specificity, pp (95% CI)	Delta PPV, pp (95% CI)
10R	EN	-1.2 (-4.8 to + 2.2)	+ 13.4 (+ 7.5 to + 19.8)	+ 10.0 (+ 5.1 to + 15.5)
10R	HGB	-0.1 (-0.9 to + 0.7)	-5.0 (-10.5 to + 0.3)	-3.4 (-7.1 to + 0.1)
10R	SVM	-1.0 (-4.6 to + 2.2)	+ 8.8 (+ 2.7 to + 15.5)	+ 6.3 (+ 1.8 to + 11.0)
10L	EN	-0.5 (-6.4 to + 5.2)	+ 0.6 (-5.3 to + 6.0)	+ 0.3 (-6.0 to + 6.0)
10L	HGB	+ 0.2 (-2.6 to + 4.1)	-0.3 (-3.1 to + 2.8)	-0.2 (-3.0 to + 3.0)
10L	SVM	-0.2 (-6.9 to + 6.1)	-4.5 (-11.6 to + 2.0)	-4.8 (-12.3 to + 1.7)
7	EN	-0.7 (-3.7 to + 2.2)	+ 3.3 (-0.6 to + 7.2)	+ 3.4 (-0.6 to + 7.8)
7	HGB	+ 0.3 (-1.5 to + 1.9)	+ 1.7 (-0.6 to + 4.2)	+ 1.8 (-0.6 to + 4.3)
7	SVM	+ 0.1 (-2.8 to + 3.0)	-2.3 (-6.3 to + 1.9)	-2.2 (-6.3 to + 1.9)
4R	EN	-0.9 (-3.2 to + 1.0)	-2.6 (-6.9 to + 1.6)	-2.1 (-5.4 to + 1.1)
4R	HGB	-0.0 (-0.9 to + 1.0)	-1.2 (-3.9 to + 1.6)	-0.8 (-2.9 to + 1.3)
4R	SVM	-0.9 (-3.4 to + 1.3)	-9.4 (-14.2 to -5.4)	-6.9 (-10.5 to -3.9)
4L	EN	-0.4 (-2.2 to + 1.2)	+ 6.5 (+ 1.5 to + 11.6)	+ 5.0 (+ 1.0 to + 9.2)
4L	HGB	+ 0.2 (-0.5 to + 1.3)	-16.7 (-22.5 to -10.9)	-10.2 (-14.6 to -6.2)
4L	SVM	+ 0.0 (-1.7 to + 1.5)	+ 1.0 (-5.2 to + 7.6)	+ 0.2 (-4.5 to + 5.3)
2R	EN	-4.3 (-13.2 to + 0.0)	+ 15.4 (+ 7.8 to + 22.6)	+ 10.2 (+ 3.0 to + 17.9)
2R	HGB	+ 0.0 (+ 0.0 to + 0.0)	+ 6.9 (+ 0.3 to + 12.9)	+ 4.7 (-0.1 to + 9.8)
2R	SVM	-2.4 (-6.0 to + 0.0)	-6.3 (-17.0 to + 4.3)	-4.8 (-11.2 to + 1.6)

Station-level estimates provide a useful view of where the operating-point shift arose, although they remain exploratory. No individual specificity or PPV comparison survived Holm adjustment across the 72-comparison family; conditional bootstrap intervals were not multiplicity-adjusted.

Retrospective comparisons at similar achieved sensitivity

To distinguish training-derived threshold effects from differences at nearly identical observed sensitivity, we also examined a retrospective held-out matched-sensitivity analysis (Tables 6 and 7). The 10R EN

differences narrowed to + 1.3 points for specificity and + 0.5 for PPV, while 4L differences changed direction. These estimates varied across stations. Because the thresholds were selected using held-out labels, this analysis is explanatory rather than a validated clinical decision strategy; it indicates that the primary finding concerns the performance of training-derived, transferred operating thresholds rather than an established advantage at exactly equal sensitivity.

Table 6
Retrospective matched-sensitivity specificity analysis. Thresholds use held-out outcomes and are explanatory rather than independently validated operating rules. Ties can prevent exact sensitivity matching. Intervals are conditional on retained predictions.

Station	Model	Empirical sensitivity %	SUV_LR specificity %	Context specificity %	Delta specificity, pp (95% CI)
10R	EN	95.7	74.5	75.8	+ 1.3 (-32.1 to + 41.1)
10R	HGB	95.7	74.5	64.3	-10.1 (-18.1 to + 4.6)
10R	SVM	95.7	74.5	77.1	+ 2.6 (-13.2 to + 46.4)
10L	EN	95.2	90.0	87.1	-2.9 (-30.5 to + 12.7)
10L	HGB	95.7	90.0	87.1	-2.9 (-8.3 to + 3.4)
10L	SVM	95.2	90.0	83.5	-6.5 (-25.2 to + 9.3)
7	EN	95.2	86.7	90.0	+ 3.3 (-6.2 to + 17.3)
7	HGB	95.2	86.7	89.6	+ 2.8 (-3.3 to + 15.5)
7	SVM	95.2	86.7	86.7	-0.1 (-12.7 to + 14.4)
4R	EN	95.7	84.8	70.3	-14.5 (-22.0 to + 4.8)
4R	HGB	95.7	84.8	81.4	-3.4 (-7.6 to + 11.1)
4R	SVM	95.7	84.8	68.4	-16.3 (-36.6 to + 4.8)
4L	EN	95.6	87.5	72.2	-15.3 (-19.8 to + 0.2)
4L	HGB	95.6	87.5	60.5	-27.0 (-30.5 to -0.3)
4L	SVM	95.6	87.5	71.0	-16.5 (-26.4 to + 6.5)
2R	EN	95.2	91.4	58.3	-33.1 (-49.4 to -2.1)
2R	HGB	95.2	91.4	88.6	-2.9 (-60.0 to -0.3)
2R	SVM	95.2	91.4	51.1	-40.3 (-66.7 to -2.1)

Table 7
Retrospective matched-sensitivity PPV analysis. Thresholds use held-out outcomes and are explanatory rather than independently validated operating rules. Intervals are conditional on retained predictions.

Station	Model	SUV_LR PPV %	Context PPV %	Delta PPV, pp (95% CI)
10R	EN	74.6	75.1	+ 0.5 (-18.8 to + 21.9)
10R	HGB	74.6	67.8	-6.8 (-12.4 to + 1.9)
10R	SVM	74.6	76.0	+ 1.4 (-10.0 to + 25.3)
10L	EN	85.9	82.6	-3.3 (-22.0 to + 12.4)
10L	HGB	85.9	83.1	-2.8 (-6.6 to + 3.9)
10L	SVM	85.9	78.6	-7.3 (-23.4 to + 8.3)
7	EN	83.7	87.2	+ 3.5 (-6.0 to + 13.6)
7	HGB	83.7	86.7	+ 3.0 (-3.2 to + 12.0)
7	SVM	83.7	83.6	-0.1 (-11.3 to + 10.9)
4R	EN	85.7	75.4	-10.2 (-14.6 to + 3.4)
4R	HGB	85.7	83.1	-2.6 (-5.4 to + 7.5)
4R	SVM	85.7	74.3	-11.3 (-20.5 to + 2.9)
4L	EN	76.7	62.0	-14.7 (-19.7 to + 1.3)
4L	HGB	76.7	56.6	-20.1 (-23.3 to -0.2)
4L	SVM	76.7	59.2	-17.5 (-27.7 to + 3.9)
2R	EN	87.1	60.8	-26.3 (-34.7 to -2.7)
2R	HGB	87.1	83.6	-3.5 (-35.2 to -0.2)
2R	SVM	87.1	56.4	-30.7 (-40.9 to -2.0)

Discrimination, calibration, and decision-curve assessment

Changes in discrimination and probability accuracy depended on station and model (Table 8). The 36 AUC/Brier comparisons yielded no Holm-adjusted significant differences. Calibration estimates likewise varied across stations (Tables 9 and 10; Fig. 3), indicating that an advantageous classification threshold should be distinguished from an improvement in the calibration of individual predicted probabilities.

Table 8
Absolute discrimination and probability error. ROC-AUC and Brier score are means across repeated outer-test predictions.

Station	Model	ROC-AUC	Brier score
10R	SUV_LR	0.939	0.086
10R	EN	0.947	0.074
10R	HGB	0.931	0.087
10R	SVM	0.944	0.081
10L	SUV_LR	0.951	0.095
10L	EN	0.942	0.088
10L	HGB	0.950	0.077
10L	SVM	0.930	0.103
7	SUV_LR	0.966	0.070
7	EN	0.969	0.065
7	HGB	0.970	0.058
7	SVM	0.966	0.066
4R	SUV_LR	0.944	0.086
4R	EN	0.941	0.082
4R	HGB	0.955	0.072
4R	SVM	0.937	0.085
4L	SUV_LR	0.958	0.069
4L	EN	0.946	0.072
4L	HGB	0.935	0.069
4L	SVM	0.941	0.079
2R	SUV_LR	0.967	0.082
2R	EN	0.918	0.100
2R	HGB	0.912	0.084
2R	SVM	0.891	0.113

Table 9

Comparative calibration on retained outer predictions. Means and ranges are across ten repetitions of pooled outer-test predictions. Ranges reflect partition sensitivity and are not confidence intervals.

Station	Model	Mean CITL	Mean joint intercept	Mean slope	Slope range
10R	SUV_LR	0.003	0.049	1.068	0.917 to 1.253
10R	EN	0.063	-0.094	0.771	0.712 to 0.830
10R	HGB	-0.012	-0.034	0.903	0.828 to 0.966
10R	SVM	-0.026	-0.085	0.871	0.726 to 1.004
10L	SUV_LR	-0.017	0.051	1.094	0.845 to 1.420
10L	EN	-0.002	-0.151	0.705	0.478 to 0.914
10L	HGB	-0.013	0.007	0.888	0.707 to 1.141
10L	SVM	-0.233	-0.268	0.779	0.619 to 0.955
7	SUV_LR	-0.007	0.025	1.035	0.944 to 1.126
7	EN	0.068	-0.098	0.816	0.734 to 0.910
7	HGB	0.012	0.011	0.981	0.892 to 1.069
7	SVM	0.004	-0.033	0.949	0.887 to 1.067
4R	SUV_LR	0.005	0.031	1.050	0.924 to 1.123
4R	EN	0.040	-0.031	0.861	0.753 to 0.920
4R	HGB	0.008	0.006	0.975	0.912 to 1.056
4R	SVM	-0.020	-0.062	0.902	0.789 to 1.001
4L	SUV_LR	0.027	0.151	1.096	0.929 to 1.273
4L	EN	0.035	-0.262	0.743	0.514 to 1.022
4L	HGB	-0.100	-0.100	0.966	0.849 to 1.060
4L	SVM	-0.011	-0.039	0.965	0.824 to 1.246
2R	SUV_LR	0.006	0.623	1.679	0.938 to 2.750
2R	EN	0.396	-0.264	0.504	0.252 to 0.748
2R	HGB	-0.099	-0.091	0.889	0.741 to 1.134
2R	SVM	-0.523	-0.610	0.511	0.251 to 0.774

Table 10

SUVmax benchmark calibration with bootstrap intervals. Bootstrap intervals are conditional on retained predictions. Valid draws denote finite joint-calibration fits.

Station	CITL (95% CI)	Joint intercept	Slope (95% CI)	Valid/2000
10R	0.003 (-0.452, 0.387)	0.049 (-0.479, 1.115)	1.068 (0.703, 1.970)	1997
10L	-0.017 (-0.579, 0.427)	0.051 (-0.576, 2.118)	1.094 (0.634, 3.211)	1932
7	-0.007 (-0.377, 0.288)	0.025 (-0.466, 0.984)	1.035 (0.711, 1.767)	2000
4R	0.005 (-0.318, 0.284)	0.031 (-0.351, 0.752)	1.050 (0.691, 1.796)	1999
4L	0.027 (-0.661, 0.537)	0.151 (-0.739, 1.670)	1.096 (0.774, 1.955)	1999
2R	0.006 (-0.734, 0.574)	0.623 (-0.534, 5.054)	1.679 (1.003, 6.302)	1758

Decision-curve analysis provided an additional view of the models across illustrative probability thresholds (Table 11; Fig. 4). Net-benefit differences varied in sign and magnitude between stations and thresholds. These findings help define the clinical decision settings in which the observed operating-point shift deserves further testing.

Table 11

Paired decision-curve net-benefit differences. Contextual model minus SUV_LR at illustrative probability thresholds. Values are net-benefit units per verified station followed by conditional 95% patient-cluster bootstrap intervals.

Station	Model	Threshold 0.10	Threshold 0.20	Threshold 0.30
10R	EN	+ 0.0147 (-0.0016, + 0.0267)	+ 0.0087 (-0.0039, + 0.0259)	+ 0.0120 (-0.0032, + 0.0319)
10R	HGB	+ 0.0090 (-0.0051, + 0.0207)	-0.0113 (-0.0255, + 0.0008)	-0.0096 (-0.0275, + 0.0068)
10R	SVM	+ 0.0161 (+ 0.0046, + 0.0282)	+ 0.0000 (-0.0127, + 0.0140)	+ 0.0003 (-0.0188, + 0.0217)
10L	EN	+ 0.0207 (+ 0.0032, + 0.0343)	-0.0025 (-0.0322, + 0.0234)	-0.0122 (-0.0388, + 0.0123)
10L	HGB	+ 0.0228 (-0.0002, + 0.0396)	+ 0.0012 (-0.0127, + 0.0141)	-0.0053 (-0.0180, + 0.0070)
10L	SVM	+ 0.0163 (+ 0.0038, + 0.0269)	-0.0058 (-0.0324, + 0.0192)	-0.0197 (-0.0500, + 0.0097)
7	EN	+ 0.0073 (+ 0.0001, + 0.0145)	+ 0.0047 (-0.0088, + 0.0191)	+ 0.0026 (-0.0116, + 0.0170)
7	HGB	+ 0.0086 (+ 0.0014, + 0.0149)	+ 0.0051 (-0.0028, + 0.0133)	+ 0.0045 (-0.0070, + 0.0178)
7	SVM	+ 0.0037 (-0.0052, + 0.0116)	+ 0.0010 (-0.0132, + 0.0169)	-0.0013 (-0.0159, + 0.0145)
4R	EN	+ 0.0099 (-0.0027, + 0.0191)	-0.0052 (-0.0176, + 0.0057)	+ 0.0002 (-0.0168, + 0.0165)
4R	HGB	+ 0.0186 (+ 0.0046, + 0.0300)	+ 0.0049 (-0.0047, + 0.0138)	+ 0.0039 (-0.0066, + 0.0154)
4R	SVM	+ 0.0066 (-0.0070, + 0.0179)	-0.0125 (-0.0254, -0.0011)	-0.0177 (-0.0366, -0.0010)
4L	EN	-0.0059 (-0.0116, -0.0010)	-0.0053 (-0.0210, + 0.0083)	-0.0015 (-0.0158, + 0.0147)
4L	HGB	-0.0056 (-0.0147, + 0.0013)	-0.0057 (-0.0152, + 0.0017)	-0.0005 (-0.0159, + 0.0177)
4L	SVM	-0.0128 (-0.0220, -0.0057)	-0.0208 (-0.0445, + 0.0000)	-0.0178 (-0.0413, + 0.0064)
2R	EN	-0.0030 (-0.0534, + 0.0327)	-0.0402 (-0.0889, -0.0036)	-0.0408 (-0.0878, -0.0023)
2R	HGB	+ 0.0103 (-0.0272, + 0.0327)	-0.0170 (-0.0326, -0.0049)	-0.0051 (-0.0204, + 0.0130)

Station	Model	Threshold 0.10	Threshold 0.20	Threshold 0.30
2R	SVM	-0.0107 (-0.0502, + 0.0165)	-0.0424 (-0.0875, -0.0112)	-0.0347 (-0.0704, -0.0005)

Inflammation, partial verification, model complexity, and explanatory features

Ablation of the pneumonia/consolidation indicator and SUVmax scarcely changed contextual-model discrimination or Brier score (Table 12), and the main EN operating-point differences were similarly preserved (Table 13). The observed specificity shift therefore did not depend primarily on a highly incomplete inflammatory measurement.

Table 12

Inflammatory-variable ablation: discrimination and Brier score. The pneumonia/consolidation indicator and SUVmax were removed together. Full-context fold-specific hyperparameters were retained to isolate feature removal.

Station	Model	Full AUC	No-inflammation AUC	Delta AUC	Full Brier	No-inflammation Brier
10R	EN	0.947	0.947	+ 0.000	0.074	0.073
10R	HGB	0.931	0.932	+ 0.001	0.087	0.087
10R	SVM	0.944	0.943	-0.001	0.081	0.081
10L	EN	0.942	0.943	+ 0.001	0.088	0.086
10L	HGB	0.950	0.950	+ 0.000	0.077	0.076
10L	SVM	0.930	0.937	+ 0.006	0.103	0.095
7	EN	0.969	0.969	+ 0.000	0.065	0.065
7	HGB	0.970	0.970	-0.000	0.058	0.058
7	SVM	0.966	0.970	+ 0.003	0.066	0.064
4R	EN	0.941	0.943	+ 0.001	0.082	0.081
4R	HGB	0.955	0.956	+ 0.000	0.072	0.071
4R	SVM	0.937	0.945	+ 0.007	0.085	0.079
4L	EN	0.946	0.946	+ 0.000	0.072	0.068
4L	HGB	0.935	0.936	+ 0.001	0.069	0.069
4L	SVM	0.941	0.943	+ 0.002	0.079	0.075
2R	EN	0.918	0.931	+ 0.013	0.100	0.098
2R	HGB	0.912	0.914	+ 0.002	0.084	0.084
2R	SVM	0.891	0.903	+ 0.012	0.113	0.103

Table 13
Inflammatory-variable ablation at the 95% target. Full and no-inflammation models used the same outer/inner partitions; values are repeat-pooled means.

Station	Model	Full specificity %	No-inflammation specificity %	Full PPV %	No-inflammation PPV %
10R	EN	81.8	81.7	79.6	79.6
10R	HGB	63.5	63.4	66.3	66.4
10R	SVM	77.3	76.2	76.0	75.2
10L	EN	85.6	86.2	80.7	81.3
10L	HGB	84.7	84.7	80.2	80.2
10L	SVM	80.5	82.0	75.5	77.4
7	EN	89.8	89.8	86.9	86.9
7	HGB	88.3	88.1	85.3	85.2
7	SVM	84.3	87.9	81.2	85.0
4R	EN	79.9	81.5	81.8	82.9
4R	HGB	81.4	81.1	83.0	82.9
4R	SVM	73.1	76.6	77.0	79.4
4L	EN	72.2	71.8	59.9	59.9
4L	HGB	49.0	51.5	44.7	45.9
4L	SVM	66.7	63.2	55.2	52.8
2R	EN	76.6	78.6	70.8	72.5
2R	HGB	68.0	67.1	65.3	64.7
2R	SVM	54.9	54.6	55.8	55.7

Inverse-probability verification weighting changed estimated sensitivity and specificity to varying degrees across stations (Table 14). Extreme weights and smaller effective sample sizes highlighted the importance of how stations were selected for pathological sampling. The verification analysis is informative about this source of uncertainty but cannot establish outcomes for unverified stations. Smaller stations, particularly 2R, also had relatively few malignant events per contextual predictor (Table 15). Detailed Holm-adjusted operating-point comparisons are presented in Tables 16 and 17.

Table 14

Verification-weighting sensitivity analysis. Verified estimates are compared with cross-fitted inverse-probability weighted estimates under a missing-at-random scenario. ESS denotes effective sample size.

Station	Sensitivity verified -> MAR	Specificity verified -> MAR	Weight ESS	Max weight
10R	0.954 -> 0.971	0.685 -> 0.701	78.9	20.00
10L	0.945 -> 0.895	0.850 -> 0.910	58.4	20.00
7	0.952 -> 0.940	0.866 -> 0.881	210.7	20.00
4R	0.951 -> 0.915	0.826 -> 0.843	221.9	8.59
4L	0.951 -> 0.914	0.657 -> 0.663	74.4	20.00
2R	0.957 -> 0.904	0.611 -> 0.589	41.8	20.00

Table 15

Station sample size and nominal events per predictor. The information unit is one unique patient per station; repeated validation does not multiply n. Events/12 is descriptive and is not a formal sample-size adequacy criterion.

Station	Verified n	Malignant / benign	Prevalence %	Events / 12	Approx. training events / 12
10R	161	69 / 92	42.9	5.75	4.60
10L	108	42 / 66	38.9	3.50	2.80
7	353	147 / 206	41.6	12.25	9.80
4R	283	138 / 145	48.8	11.50	9.20
4L	150	45 / 105	30.0	3.75	3.00
2R	56	21 / 35	37.5	1.75	1.40
2L	15	0 / 15	0.0	Not modeled	Not modeled

Table 16

Multiplicity-adjusted comparisons at the 95% target. All 72 contextual-model-versus-SUV_LR specificity/PPV contrasts across both sensitivity targets form one exploratory Holm family. Values are fold-mean percentage-point difference followed by raw P / Holm-adjusted P.

Station	Model	Specificity difference; P / adjusted P	PPV difference; P / adjusted P
10R	EN	+ 13.31; 0.072 / 1.000	+ 9.35; 0.070 / 1.000
10R	HGB	-4.99; 0.633 / 1.000	-3.10; 0.632 / 1.000
10R	SVM	+ 8.61; 0.377 / 1.000	+ 5.95; 0.410 / 1.000
10L	EN	+ 0.65; 0.927 / 1.000	-0.70; 0.896 / 1.000
10L	HGB	-0.27; 0.968 / 1.000	-0.68; 0.890 / 1.000
10L	SVM	-4.51; 0.594 / 1.000	-5.33; 0.417 / 1.000
7	EN	+ 3.27; 0.239 / 1.000	+ 3.48; 0.235 / 1.000
7	HGB	+ 1.71; 0.454 / 1.000	+ 1.83; 0.439 / 1.000
7	SVM	-2.26; 0.558 / 1.000	-2.06; 0.591 / 1.000
4R	EN	-2.62; 0.623 / 1.000	-2.08; 0.561 / 1.000
4R	HGB	-1.17; 0.794 / 1.000	-0.86; 0.783 / 1.000
4R	SVM	-9.45; 0.103 / 1.000	-7.04; 0.075 / 1.000
4L	EN	+ 6.48; 0.555 / 1.000	+ 4.17; 0.604 / 1.000
4L	HGB	-16.67; 0.199 / 1.000	-10.74; 0.208 / 1.000
4L	SVM	+ 0.95; 0.939 / 1.000	-1.98; 0.820 / 1.000
2R	EN	+ 15.43; 0.450 / 1.000	+ 6.45; 0.617 / 1.000
2R	HGB	+ 6.86; 0.699 / 1.000	+ 2.02; 0.834 / 1.000
2R	SVM	-6.29; 0.776 / 1.000	-8.41; 0.520 / 1.000

Table 17

Multiplicity-adjusted comparisons at the 90% target. All 72 contextual-model-versus-SUV_LR specificity/PPV contrasts across both sensitivity targets form one exploratory Holm family. Values are fold-mean percentage-point difference followed by raw P / Holm-adjusted P.

Station	Model	Specificity difference; P / adjusted P	PPV difference; P / adjusted P
10R	EN	+ 3.59; 0.165 / 1.000	+ 4.13; 0.140 / 1.000
10R	HGB	-1.61; 0.674 / 1.000	-1.31; 0.711 / 1.000
10R	SVM	+ 0.66; 0.888 / 1.000	+ 0.51; 0.915 / 1.000
10L	EN	-2.58; 0.486 / 1.000	-3.13; 0.489 / 1.000
10L	HGB	-1.34; 0.622 / 1.000	-1.99; 0.593 / 1.000
10L	SVM	-7.73; 0.233 / 1.000	-8.32; 0.214 / 1.000
7	EN	+ 0.89; 0.684 / 1.000	+ 0.88; 0.730 / 1.000
7	HGB	+ 0.73; 0.685 / 1.000	+ 0.71; 0.726 / 1.000
7	SVM	-0.33; 0.895 / 1.000	-0.57; 0.843 / 1.000
4R	EN	-0.83; 0.616 / 1.000	-0.91; 0.578 / 1.000
4R	HGB	+ 0.41; 0.807 / 1.000	+ 0.43; 0.798 / 1.000
4R	SVM	-3.52; 0.202 / 1.000	-3.35; 0.178 / 1.000
4L	EN	-1.43; 0.686 / 1.000	-1.78; 0.702 / 1.000
4L	HGB	-4.57; 0.482 / 1.000	-4.94; 0.512 / 1.000
4L	SVM	-9.24; 0.113 / 1.000	-11.97; 0.072 / 1.000
2R	EN	-5.14; 0.432 / 1.000	-4.81; 0.503 / 1.000
2R	HGB	-7.71; 0.409 / 1.000	-6.68; 0.353 / 1.000
2R	SVM	-18.00; 0.095 / 1.000	-17.01; 0.096 / 1.000

Held-out grouped permutation analyses implicated uptake in other nodal stations and primary-tumor measurements in some models, consistent with the rationale for patient-specific metabolic context (Tables 18–20). The relative contributions varied by model and station, making the signal more consistent with a distributed metabolic pattern than with a single dominant supplementary variable.

Table 18
EN grouped held-out permutation importance. Mean held-out AUC decrease after jointly shuffling each contextual group. Approximate intervals use dependence-corrected fold-level variation and are exploratory.

Station	Contextual group	AUC decrease	Approximate 95% interval
10R	Other nodal stations	+ 0.0003	-0.0079 to + 0.0085
10R	Aortic blood pool	-0.0002	-0.0018 to + 0.0015
10R	Primary tumor	+ 0.0264	+ 0.0074 to + 0.0455
10R	Laterality	-0.0001	-0.0106 to + 0.0104
10R	Inflammatory variables	-0.0001	-0.0013 to + 0.0010
10L	Other nodal stations	+ 0.0562	+ 0.0105 to + 0.1019
10L	Aortic blood pool	+ 0.0007	-0.0088 to + 0.0101
10L	Primary tumor	+ 0.0017	-0.0051 to + 0.0084
10L	Laterality	+ 0.0197	-0.0064 to + 0.0457
10L	Inflammatory variables	-0.0006	-0.0051 to + 0.0039
7	Other nodal stations	+ 0.0200	+ 0.0094 to + 0.0306
7	Aortic blood pool	-0.0003	-0.0012 to + 0.0006
7	Primary tumor	+ 0.0338	+ 0.0204 to + 0.0471
7	Laterality	-0.0001	-0.0007 to + 0.0005
7	Inflammatory variables	-0.0001	-0.0006 to + 0.0004
4R	Other nodal stations	+ 0.0305	+ 0.0141 to + 0.0470
4R	Aortic blood pool	+ 0.0001	-0.0007 to + 0.0010
4R	Primary tumor	+ 0.0141	+ 0.0045 to + 0.0238
4R	Laterality	-0.0007	-0.0029 to + 0.0016
4R	Inflammatory variables	+ 0.0007	-0.0051 to + 0.0065
4L	Other nodal stations	-0.0021	-0.0159 to + 0.0118
4L	Aortic blood pool	+ 0.0016	-0.0036 to + 0.0068
4L	Primary tumor	+ 0.0087	-0.0066 to + 0.0240
4L	Laterality	+ 0.0011	-0.0106 to + 0.0128
4L	Inflammatory variables	+ 0.0003	-0.0083 to + 0.0089
2R	Other nodal stations	+ 0.1170	+ 0.0346 to + 0.1995

Station	Contextual group	AUC decrease	Approximate 95% interval
2R	Aortic blood pool	-0.0032	-0.0281 to + 0.0218
2R	Primary tumor	-0.0065	-0.0376 to + 0.0246
2R	Laterality	-0.0001	-0.0024 to + 0.0022
2R	Inflammatory variables	-0.0030	-0.0198 to + 0.0138

Table 19
HGB grouped held-out permutation importance. Mean held-out AUC decrease after jointly shuffling each contextual group. Approximate intervals use dependence-corrected fold-level variation and are exploratory.

Station	Contextual group	AUC decrease	Approximate 95% interval
10R	Other nodal stations	-0.0023	-0.0115 to + 0.0069
10R	Aortic blood pool	+ 0.0024	-0.0070 to + 0.0117
10R	Primary tumor	+ 0.0099	-0.0042 to + 0.0240
10R	Laterality	+ 0.0002	-0.0007 to + 0.0010
10R	Inflammatory variables	-0.0005	-0.0024 to + 0.0015
10L	Other nodal stations	+ 0.0180	-0.0067 to + 0.0427
10L	Aortic blood pool	+ 0.0123	-0.0090 to + 0.0336
10L	Primary tumor	-0.0004	-0.0086 to + 0.0078
10L	Laterality	+ 0.0000	+ 0.0000 to + 0.0000
10L	Inflammatory variables	-0.0003	-0.0019 to + 0.0013
7	Other nodal stations	+ 0.0079	+ 0.0007 to + 0.0150
7	Aortic blood pool	+ 0.0004	-0.0031 to + 0.0039
7	Primary tumor	+ 0.0125	+ 0.0021 to + 0.0229
7	Laterality	-0.0000	-0.0000 to + 0.0000
7	Inflammatory variables	-0.0002	-0.0016 to + 0.0013
4R	Other nodal stations	+ 0.0125	-0.0005 to + 0.0255
4R	Aortic blood pool	+ 0.0091	+ 0.0010 to + 0.0172
4R	Primary tumor	+ 0.0072	-0.0016 to + 0.0161
4R	Laterality	-0.0000	-0.0001 to + 0.0000
4R	Inflammatory variables	+ 0.0004	-0.0049 to + 0.0057
4L	Other nodal stations	+ 0.0063	-0.0150 to + 0.0276
4L	Aortic blood pool	+ 0.0019	-0.0057 to + 0.0094
4L	Primary tumor	+ 0.0037	-0.0100 to + 0.0174
4L	Laterality	+ 0.0000	-0.0002 to + 0.0002
4L	Inflammatory variables	-0.0014	-0.0059 to + 0.0031
2R	Other nodal stations	-0.0018	-0.0457 to + 0.0421

Station	Contextual group	AUC decrease	Approximate 95% interval
2R	Aortic blood pool	+ 0.0042	-0.0170 to + 0.0254
2R	Primary tumor	-0.0020	-0.0097 to + 0.0057
2R	Laterality	+ 0.0000	+ 0.0000 to + 0.0000
2R	Inflammatory variables	-0.0033	-0.0172 to + 0.0106

Table 20
SVM grouped held-out permutation importance. Mean held-out AUC decrease after jointly shuffling each contextual group. Approximate intervals use dependence-corrected fold-level variation and are exploratory.

Station	Contextual group	AUC decrease	Approximate 95% interval
10R	Other nodal stations	+ 0.0213	+ 0.0040 to + 0.0385
10R	Aortic blood pool	+ 0.0020	-0.0042 to + 0.0081
10R	Primary tumor	+ 0.0243	+ 0.0077 to + 0.0410
10R	Laterality	+ 0.0008	-0.0054 to + 0.0069
10R	Inflammatory variables	+ 0.0058	-0.0042 to + 0.0159
10L	Other nodal stations	+ 0.0912	+ 0.0310 to + 0.1515
10L	Aortic blood pool	-0.0023	-0.0134 to + 0.0088
10L	Primary tumor	+ 0.0025	-0.0053 to + 0.0103
10L	Laterality	+ 0.0254	+ 0.0026 to + 0.0483
10L	Inflammatory variables	-0.0020	-0.0116 to + 0.0076
7	Other nodal stations	+ 0.0235	+ 0.0105 to + 0.0366
7	Aortic blood pool	+ 0.0006	-0.0031 to + 0.0042
7	Primary tumor	+ 0.0306	+ 0.0163 to + 0.0448
7	Laterality	-0.0001	-0.0030 to + 0.0028
7	Inflammatory variables	-0.0002	-0.0037 to + 0.0033
4R	Other nodal stations	+ 0.0325	+ 0.0077 to + 0.0574
4R	Aortic blood pool	-0.0044	-0.0121 to + 0.0034
4R	Primary tumor	+ 0.0130	+ 0.0033 to + 0.0227
4R	Laterality	-0.0006	-0.0036 to + 0.0024
4R	Inflammatory variables	-0.0020	-0.0111 to + 0.0072
4L	Other nodal stations	-0.0016	-0.0224 to + 0.0191
4L	Aortic blood pool	+ 0.0054	-0.0077 to + 0.0184
4L	Primary tumor	+ 0.0138	+ 0.0001 to + 0.0275
4L	Laterality	+ 0.0020	-0.0078 to + 0.0118
4L	Inflammatory variables	+ 0.0074	-0.0095 to + 0.0242
2R	Other nodal stations	+ 0.1823	+ 0.0931 to + 0.2715

Station	Contextual group	AUC decrease	Approximate 95% interval
2R	Aortic blood pool	-0.0007	-0.0306 to + 0.0292
2R	Primary tumor	-0.0073	-0.0330 to + 0.0184
2R	Laterality	-0.0035	-0.0148 to + 0.0078
2R	Inflammatory variables	+ 0.0009	-0.0370 to + 0.0388

Discussion

The main finding is clinically focused: at a stringent high-sensitivity operating point, patient-specific intrathoracic metabolic context improved the specificity and positive predictive value of station-level NSCLC PET/CT classification compared with local SUVmax alone. This was observed against an already strong SUVmax benchmark, using pathology-confirmed outcomes and thresholds derived independently of held-out patients. The benefit lies in refining which FDG-avid stations are classified as suspicious, rather than simply increasing model complexity or maximizing AUC.

The magnitude of this shift is clinically interpretable. At the verified-station disease prevalence in our cohort, scaling the observed cross-validated rate differences to a hypothetical set of 1,000 pathology-verified stations yields approximately 25 fewer false-positive classifications and four additional false-negative classifications. This is an arithmetic illustration, not a measured change in patient management. It nevertheless shows why specificity and sensitivity should be examined together: even a moderate gain in specificity can change the number of benign stations labeled suspicious, while the small associated sensitivity difference remains clinically important.

There is a plausible imaging basis for this result. The meaning of a target nodal SUVmax may depend on the surrounding metabolic landscape, including uptake in other nodal stations, the primary tumor, and mediastinal blood-pool activity. A node with a given SUVmax need not carry the same risk in a metabolically quiet thorax as in one showing widespread avidity. All contextual inputs in this study came from standard PET/CT assessment, without additional acquisition or dedicated radiomics segmentation. The feature-permutation results support investigating a distributed metabolic signal rather than assuming that one extra measurement explains the effect.

The potential application is a context-aware layer of support for nodal interpretation when high sensitivity is a priority. Rather than replacing the local SUVmax measurement, metabolic context adds information from the entire thorax to a station-level judgment. Its practical appeal is that these measurements are routinely accessible and the observed effect directly concerns the false-positive burden of a sensitive classification policy. Whether this can help prioritize tissue verification or alter clinical pathways should be assessed prospectively while preserving current pathological staging standards.

Previous PET/CT prediction studies have highlighted the promise of nodal SUVs, radiomics, machine-learning N-stage assessment, multimodality approaches, and automated node detection [8–18]. Our contribution is complementary: a pathology-verified, station-specific assessment of routinely available patient-wide metabolic information against a stringent local SUVmax benchmark, with leakage-controlled repeated validation and sensitivity-based operating rules. The result identifies a concrete setting in which added metabolic context is worth testing for clinical utility.

Sensitivity analyses also sharpen the interpretation. The specificity gain was concentrated at the 95% training-derived target; it was not reproduced at 90%, and retrospective sensitivity-matched comparisons did not show a consistent advantage. The station-specific comparisons did not survive multiplicity correction, while AUC, calibration, and decision-curve results were heterogeneous. These findings point to a threshold-dependent opportunity rather than a general improvement in every metric. They also emphasize that the operating policy, not merely the underlying ranking model, is central to the observed benefit.

Several aspects of the study warrant consideration. This was a retrospective single-center analysis using one PET/CT platform and consensus image readings rather than independent-reader measurements. Although every modeled label was pathology-confirmed, sampling was selective and an adequately negative EBUS-TBNA result is not identical to a negative surgical specimen [2, 3, 5, 25, 26]. The global EN comparison was selected post hoc, and its bootstrap intervals were conditional on stored predictions rather than full model refitting. Event counts were limited at some stations, inflammatory SUVmax was frequently missing, and external or temporal validation was not available.

The next step is an independent, prospectively specified evaluation of the contextual model and its training-derived thresholds, with explicit sensitivity safeguards and patient-clustered reporting of specificity, PPV, calibration, and clinical net benefit. A paired review of stations whose classifications change would be especially informative for identifying the imaging patterns that drive improvements. Ultimately, studies linking these changes to tissue-sampling choices and staging decisions will show whether context-aware PET/CT assessment improves the efficiency of the NSCLC diagnostic pathway. Validation sample sizes should account for patient-level dependence and event prevalence [29].

Conclusions

Patient-specific intrathoracic metabolic context increased specificity and positive predictive value in pathology-verified NSCLC nodal classification at a stringent, training-derived high-sensitivity operating point. The findings establish a clinically relevant direction for refining local SUVmax assessment using information already available from routine PET/CT. Independent validation of the operating strategy is the essential next step toward clinical application.

Abbreviations

AUC

area under the receiver operating characteristic curve
CI
confidence interval
CITL
calibration-in-the-large
DCA
decision-curve analysis
EBUS-TBNA
endobronchial ultrasound-guided transbronchial needle aspiration
EN
elastic-net logistic regression
FDG
fluorodeoxyglucose
HGB
histogram gradient boosting
NPV
negative predictive value
NSCLC
non-small cell lung cancer
PET/CT
positron-emission tomography/computed tomography
PPV
positive predictive value
ROC-AUC
area under the receiver operating characteristic curve
SUV
standardized uptake value
SUVmax
maximum standardized uptake value
SUV_LR
station-specific SUVmax-only logistic regression
SVM
support-vector machine
VOI
volume of interest.

Declarations

Ethics approval and consent to participate

The study protocol was approved by the Non-Interventional Clinical Research Ethics Committee of Istanbul Yedikule Chest Diseases and Thoracic Surgery Training and Research Hospital (decision no. 2026-10, dated 26 February 2026). Informed consent was obtained from all participants. The study was conducted in accordance with the Declaration of Helsinki and the applicable institutional guidelines and regulations.

Consent for publication

Not applicable. The manuscript does not contain identifiable individual participant information, images, or videos.

Competing interests

The authors declare that they have no competing interests.

Authors' information

Not applicable.

Funding

Not applicable.

Author Contribution

R.A.: conceptualization, methodology, investigation, data curation, writing - original draft, writing - review and editing, supervision, and project administration. M.M.A.: writing - review and editing, supervision, and project administration. F.A.G.: methodology and investigation. D.T.: data curation. I.Y.: methodology, formal analysis, software, validation, writing - original draft, writing - review and editing, and supervision. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

The datasets generated and/or analysed during the current study are not publicly available because they contain patient-level clinical imaging and pathology information subject to institutional and ethical

restrictions. De-identified data may be available from the corresponding author on reasonable request, subject to approval by the responsible institutional and ethics bodies and applicable data-use requirements. Analysis code and aggregate result tables are available from the corresponding author on reasonable request.

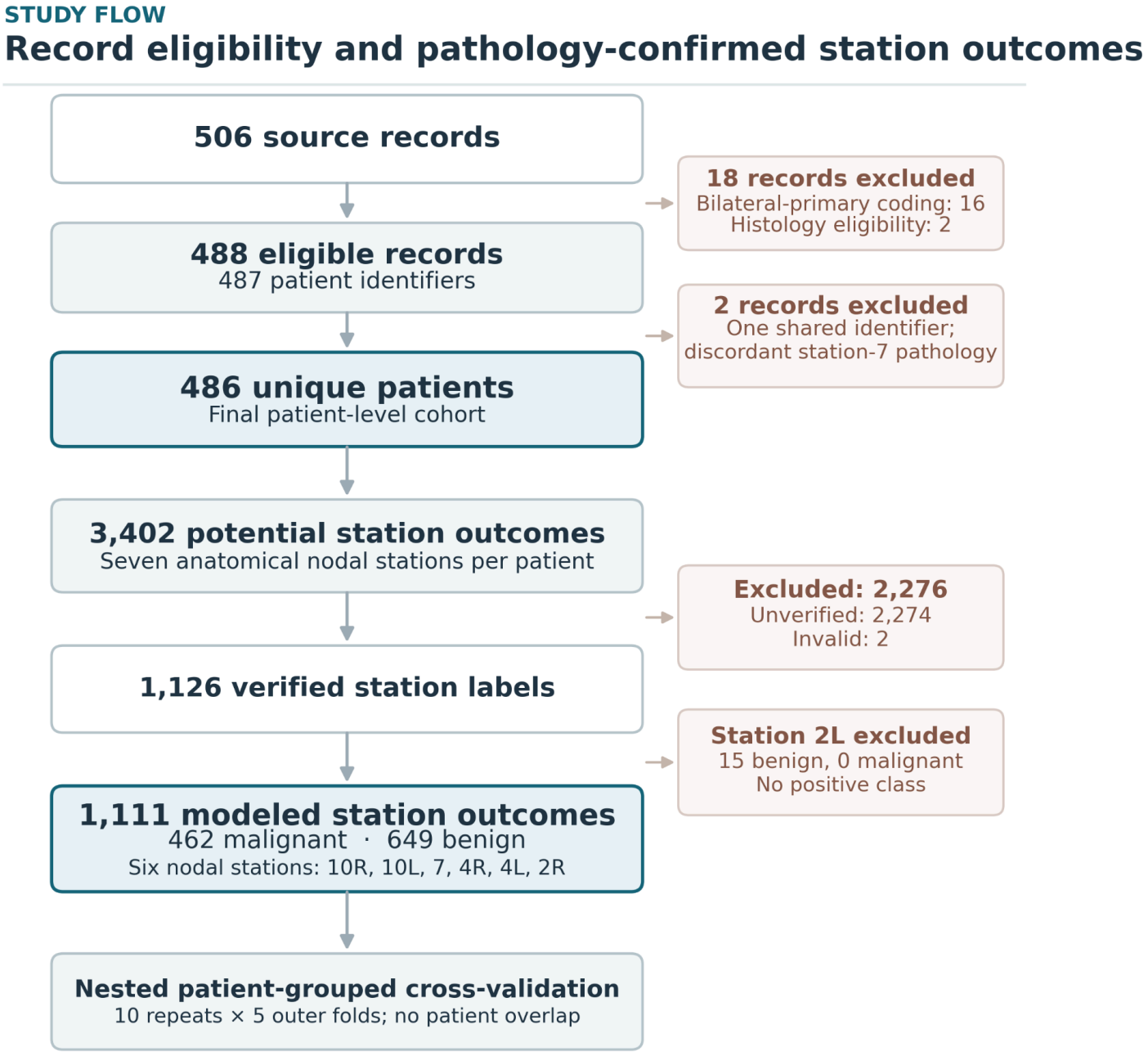
References

1. El-Sherief AH, Lau CT, Wu CC, et al. International Association for the Study of Lung Cancer (IASLC) lymph node map: a radiologic atlas and review. Radiographics. 2014;34(6):1680–91.
2. Silvestri GA, Gonzalez AV, Jantz MA, et al. Methods for staging non-small cell lung cancer: diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest. 2013;143(5 Suppl):eS211–50.
3. Schmidt-Hansen M, Baldwin DR, Zamora J. FDG-PET/CT imaging for mediastinal staging in patients with potentially resectable non-small cell lung cancer. JAMA. 2015;313(14):1465–6.
4. Mingels C, Madani MH, Sen F, et al. Diagnostic accuracy in NSCLC lymph node staging with total-body and conventional PET/CT. Eur J Nucl Med Mol Imaging. 2025;52(9):3190–8.
5. Gu P, Zhao YZ, Jiang LY, et al. Endobronchial ultrasound-guided transbronchial needle aspiration for staging of lung cancer: a systematic review and meta-analysis. Eur J Cancer. 2009;45(8):1389–96.
6. Rahman WT, Wale DJ, Viglianti BL, et al. The impact of infection and inflammation in oncologic 18F-FDG PET/CT imaging. Eur J Nucl Med Mol Imaging. 2019;46(7):1380–92.
7. Deppen SA, Blume JD, Aldrich MC, et al. Accuracy of FDG-PET to diagnose lung cancer in areas with infectious lung disease: a meta-analysis. JAMA. 2014;312(12):1227–36.
8. Yu X, Wang J, Huang L, Xie L, Su Y. Predictive value of 18F-FDG PET/CT metabolic parameters for lymph node metastasis of non-small cell lung cancer. Biomark Med. 2025;19:35–41. 10.1080/17520363.2024.2443379.
9. Liao X, Liu M, Li S, et al. The value on SUV-derived parameters assessed on 18F-FDG PET/CT for predicting mediastinal lymph node metastasis in non-small cell lung cancer. BMC Med Imaging. 2023;23:49. 10.1186/s12880-023-01004-7.
10. Wang H, Zhou Z, Li Y, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images. EJNMMI Res. 2017;7:11. 10.1186/s13550-017-0260-9.
11. Erdogdu E, Oksuz I, Duman S, et al. Machine learning-based prediction of N2 lymph node metastasis in non-small cell lung cancer. BMC Pulm Med. 2025;25(1):454. 10.1186/s12890-025-03921-5.
12. Sasaki A, Shiiya H, Ujiie H, Hirata K, Ogawa T, Narita S, Otsuka S, Yamasaki H, Ohtaka K, Fujiwara-Kuroda A, Aragaki M, Kato T. FDG-PET/CT data enhances machine learning prediction of occult lymph-node metastasis in lung adenocarcinoma. J Thorac Dis. 2026;18(5):507. 10.21037/jtd-2026-1-0339.

13. Wumener X, Hu Z, Zhao J, Wang H, Zhang Y, Deng Y, Zhao J, Liang Y. Predictive value of radiomics modeling based on dynamic and static 18F-FDG PET/CT imaging for the differential diagnosis of lymph nodes in lung cancer. Front Oncol. 2025;15:1637366. 10.3389/fonc.2025.1637366.
14. Rogasch JMM, Michaels L, Baumgartner GL, et al. A machine learning tool to improve prediction of mediastinal lymph node metastases in non-small cell lung cancer using routinely obtainable [18F]FDG-PET/CT parameters. Eur J Nucl Med Mol Imaging. 2023;50(7):2140–51. 10.1007/s00259-023-06145-z.
15. Rosian AS, Pusztai AM, Constantinescu A, Rus GA, Oancea C, Manolescu D. Artificial Intelligence-Based 18F-FDG PET/CT Radiomics for Mediastinal Lymph Node Staging in Non-Small Cell Lung Cancer: A Systematic Review. Diagnostics. 2026;16(13):2014. 10.3390/diagnostics16132014.
16. Zhai W, Li X, Zhou T, et al. A machine learning-based 18F-FDG PET/CT multi-modality fusion radiomics model to predict mediastinal-hilar lymph node metastasis in NSCLC: a multi-centre study. Clin Radiol. 2025;83:106832.
17. Wallis D, Soussan M, Lacroix M, et al. An 18F-FDG PET/CT deep learning method for fully automated detection of pathological mediastinal lymph nodes in lung cancer patients. Eur J Nucl Med Mol Imaging. 2022;49:881–8.
18. Hatt M, Tixier F, Pierce L, et al. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume. J Nucl Med. 2015;56(1):38–44.
19. Boellaard R, Delgado-Bolton R, Oyen WJG, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging. 2015;42:328–54. 10.1007/s00259-014-2961-x.
20. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. 10.1186/1471-2105-7-91.
21. Nadeau C, Bengio Y. Inference for the generalization error. Mach Learn. 2003;52:239–81. 10.1023/A:1024068626366.
22. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65–70.
23. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. 10.1186/s12916-019-1466-7.
24. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18. 10.1186/s41512-019-0064-7.
25. Kohn MA. Studies of diagnostic test accuracy: partial verification bias and test result-based sampling. J Clin Epidemiol. 2022;145:179–82. 10.1016/j.jclinepi.2022.01.022.
26. Arifin WN, Yusof UK. Correcting for partial verification bias in diagnostic accuracy studies: a tutorial using R. Stat Med. 2022;41:1709–27. 10.1002/sim.9311.
27. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.

28. El Naqa I, Boone JM, Benedict SH, et al. AI in medical physics: guidelines for publication. Med Phys. 2021;48:4711–4. 10.1002/mp.15170.
29. Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med. 2021;40:4230–51. 10.1002/sim.9025.

Figures



Only pathology-confirmed station outcomes were modeled; unverified stations were not presumed benign.

Figure 1

Patient and station-level selection flow. Of 506 source records, 18 were excluded by bilateral-primary or histology rules, and two further nonidentical records with a shared identifier and discordant station-7 pathology were excluded. The final cohort contained 486 unique patients and 3,402 potentially observable station labels. After excluding 2,274 unverified and two invalid station outcomes, 1,126 pathology-verified labels remained. Station 2L (15 benign, no malignant findings) was not modeled; the six remaining stations contributed 1,111 binary outcomes (462 malignant and 649 benign). Patient-grouped nested cross-validation prevented within-split patient overlap.

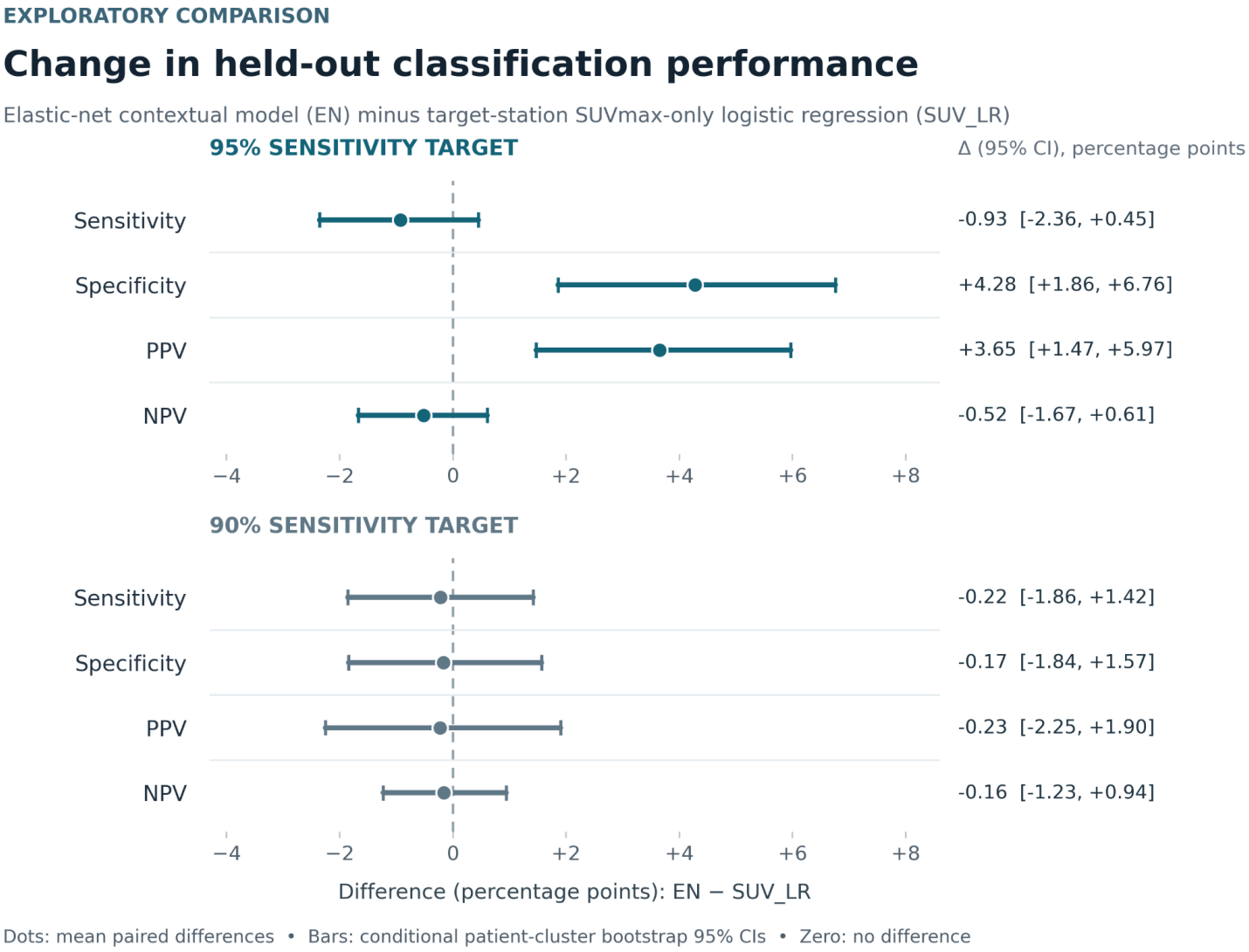


Figure 2

Micro-averaged held-out operating-point differences (contextual elastic-net logistic regression minus local SUVmax-only logistic regression) at the 95% and 90% training-sensitivity targets. Dots denote mean paired differences in percentage points; whiskers are conditional 95% patient-cluster bootstrap confidence intervals around retained cross-validated predictions. Thresholds were selected from training

data and transferred unchanged to held-out folds. The vertical dashed line indicates no difference. The pooled comparison was defined post hoc; the intervals do not incorporate complete model-development or model-selection uncertainty. PPV, positive predictive value; NPV, negative predictive value.

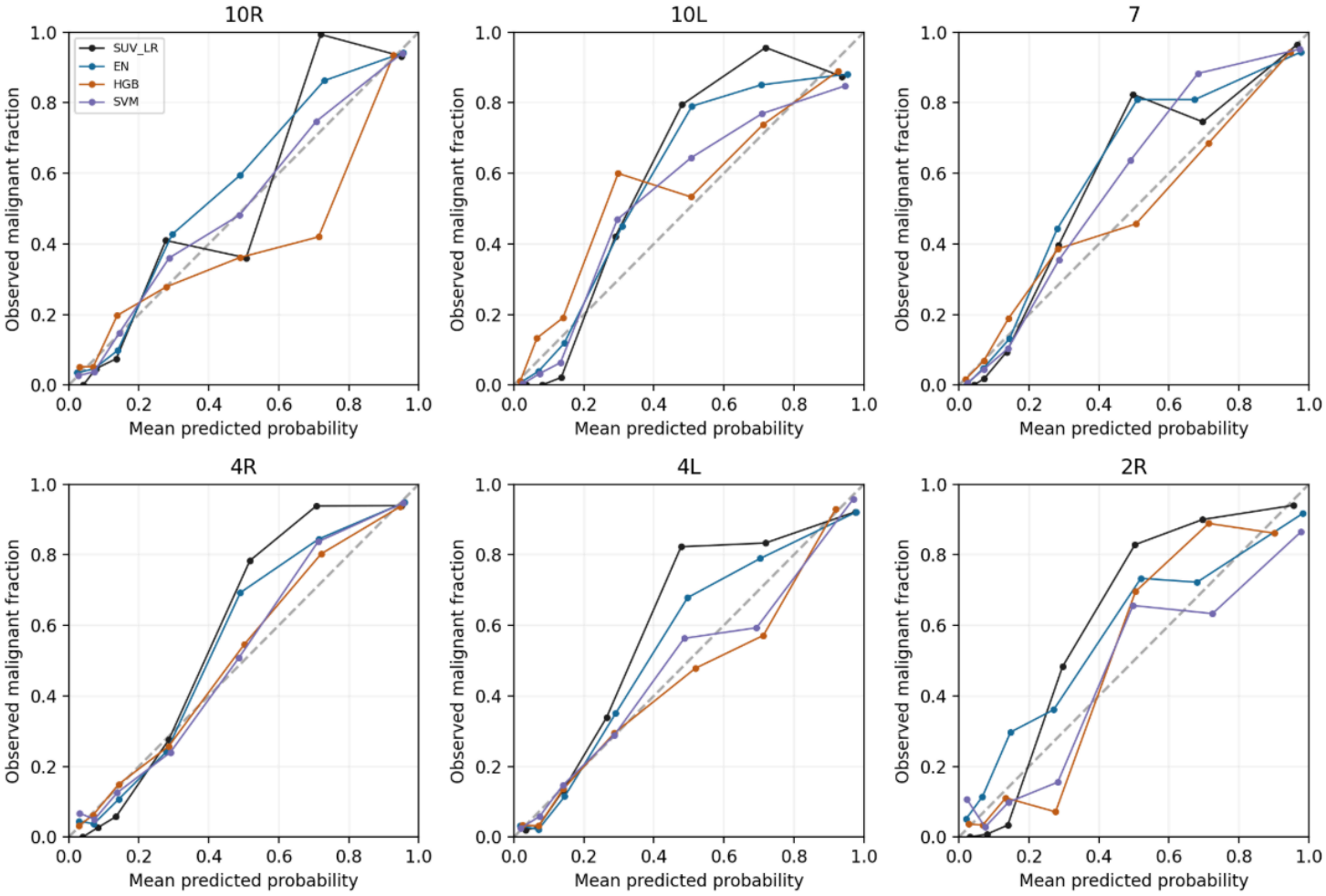


Figure 3

Calibration plots for station-specific models using retained outer-fold predictions. Observed malignancy frequency is compared with mean predicted probability; the diagonal denotes perfect calibration.

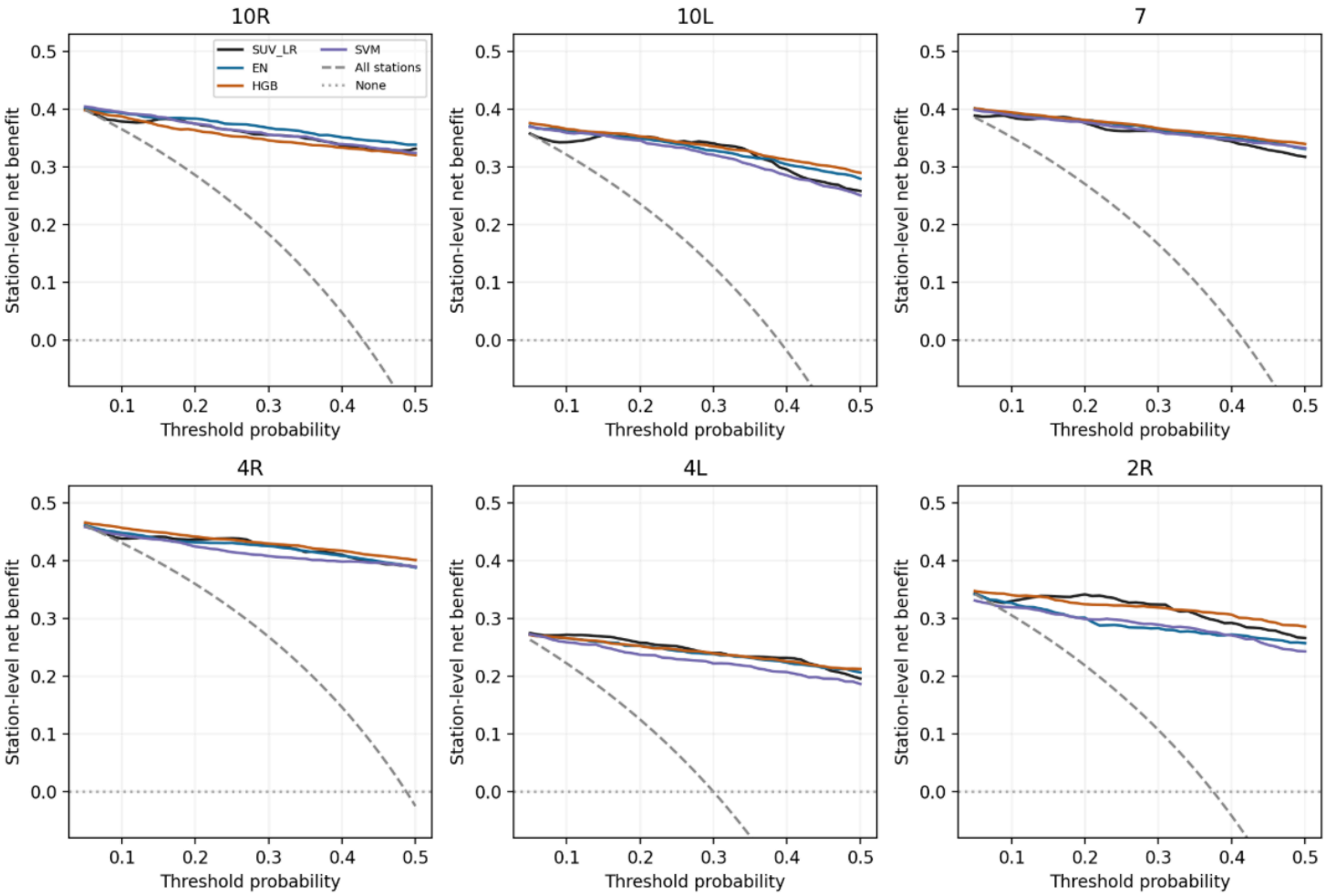


Figure 4

Decision curves for the modeled stations. Net benefit is calculated from held-out predictions across illustrative threshold probabilities; plots do not validate a clinical management policy.