Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Performance of Lung Cancer Prediction Models for Screening-detected, Incidental, and Biopsied Pulmonary Nodules

Thomas Z. Li, BS[1,2] • *Kaiwen Xu, BS, MSc*[3] • *Aravind Krishnan, BE*[4] • *Riqiang Gao, PhD*[5] • *Michael N. Kammer, PhD*[6] • *Sanja Antic, MD*[6] • *David Xiao, MD*[7] • *Michael Knight, MD*[6] • *Yency Martinez, MD*[6] • *Rafael Paez, MD, MSCI*[6] • *Robert J. Lentz, MD*[6] • *Stephen Deppen, PhD*[7] • *Eric L. Grogan, MD, MPH*[7] • *Thomas A. Lasko, MD, PhD*[3,8] • *Kim L. Sandler, MD*[9] • *Fabien Maldonado, MD, MSc*[6] • *Bennett A. Landman, PhD*[2,3,4,9]

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

See also commentary by Shao and Niu in this issue.

Radiology: Artificial Intelligence 2025; 7(2):e230506 • https://doi.org/10.1148/ryai.230506 • Content codes: AI CH CT OI

Purpose: To evaluate the performance of eight lung cancer prediction models on patient cohorts with screening-detected, incidentally detected, and bronchoscopically biopsied pulmonary nodules.

Materials and Methods: This study retrospectively evaluated promising predictive models for lung cancer prediction in three clinical settings: lung cancer screening with low-dose CT, incidentally detected pulmonary nodules, and nodules deemed suspicious enough to warrant a biopsy. The area under the receiver operating characteristic curve of eight validated models, including logistic regressions on clinical variables and radiologist nodule characterizations, artificial intelligence (AI) on chest CT scans, longitudinal imaging AI, and multimodal approaches for prediction of lung cancer risk was assessed in nine cohorts (n = 898, 896, 882, 219, 364, 117, 131, 115, 373) from multiple institutions. Each model was implemented from their published literature, and each cohort was curated from primary data sources collected over periods from 2002 to 2021.

Results: No single predictive model emerged as the highest-performing model across all cohorts, but certain models performed better in specific clinical contexts. Single-time-point chest CT AI performed well for screening-detected nodules but did not generalize well to other clinical settings. Longitudinal imaging and multimodal models demonstrated comparatively good performance on incidentally detected nodules. When applied to biopsied nodules, all models showed low performance.

Conclusion: Eight lung cancer prediction models failed to generalize well across clinical settings and sites outside of their training distributions.

Supplemental material is available for this article.

Every year, an estimated 1.57 million Americans have at least one pulmonary nodule detected either incidentally at routine chest CT or during lung cancer screening (1). Although biopsy of the nodule remains the reference standard diagnostic test for malignancy, it involves an invasive procedure associated with morbidity, mortality, additional health care costs, and anxiety for patients (2,3). With 95% of indeterminate pulmonary nodules found to be benign (4), clinical guidelines (1,5–7) recommend risk-stratifying nodules before resorting to invasive diagnostics and surgical intervention. Statistical models for predicting lung cancer have the potential to improve this risk stratification, aiding in earlier diagnosis of malignancy as well as reducing morbidity, costs, and anxiety associated with the workup of benign disease.

Validated predictive models developed to stratify pulmonary nodules consist of clinical prediction models, cross-sectional or longitudinal artificial intelligence (AI) models, and multimodal approaches. We consider a predictive model validated if it has demonstrated competitive discriminatory performance (area under the receiver operating characteristic curve [AUC] above 0.75) on a separate test cohort. The Brock (8) and Mayo (9) models are two of the most used models in clinical practice and recommended by clinical guidelines. They are well validated logistic regressions and are based on readily available variables, such as demographics (10), smoking history, and radiologist assessment of nodule features. However, they require radiologists to first detect and characterize the nodule, a step that can be subject to interreader variability (11–13).

Recent research has validated several AI models for cancer prediction. These operate directly on the voxels of the image, negating the need for radiologists to first describe nodule morphology or measure sizes. One of the early AI successes in lung cancer prediction was the study by Liao et al (14). Their two-step approach involved first detecting suspicious lesions in the lung field from a single chest CT image and then computing malignancy risk from the proposed regions of interest. Recently, Mikhael et al (15) publicly released Sybil, a predictive model that extracts global chest features along with regional attention features to predict lung cancer risk up to 6 years.

Previous work has also leveraged AI on longitudinal imaging. Gao et al (16) and Li et al (17) extended the work of Liao et al (14) to leverage consecutive chest CT scans and the time between scans. In another longitudinal imaging approach, Ardila et al (18) demonstrated impressive AUC performance of a model including global chest features outside of local regions of interest, but their model was not released publicly. Recently, efforts that leveraged data from multiple modalities have shown (19,20), with limited validation, that the combination of clinical variables and imaging AI can improve performance over single-modality approaches.

Abbreviations

AUC = area under the receiver operating characteristic curve, AI = artificial intelligence, DECAMP = Detection of Early Lung Cancer Among Military Personnel, LI-VUMC = Longitudinal Incidental-VUMC, MCL = Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions, NSCLC = non–small cell lung cancer, NLST = National Lung Screening Trial, SCLC = small cell lung cancer, UCD = University of Colorado Denver, UPMC = University of Pittsburgh Medical Center, VLSP = Vanderbilt Lung Screening Program, VUMC = Vanderbilt University Medical Center

Summary

The performance of eight different statistical models for lung cancer prediction depended heavily on the clinical setting they were applied in, and models generalized poorly on sites outside of their training distributions.

Key Points

- Models predicting lung cancer risk from a single time point had a higher area under the receiver operating characteristic curve (AUC) for patients who underwent lung cancer screening but performed worse in patients with incidentally detected nodules relative to other models.
- Longitudinal models and multimodal models had comparatively high AUCs for patients with incidentally detected nodules but showed lower performance than single-time-point models in lung cancer screening cohorts.
- Both clinical variable–based models and artificial intelligence–based models performed poorly in patients with pulmonary nodules who were triaged for invasive biopsies.

Keywords

Diagnosis, Classification, Application Domain, Lung

The plethora of research is promising, but several concerns arise when considering the clinical utility of predictive models in lung cancer diagnostics. First, the comparative advantage of AI models versus commonly used linear models has not been quantitatively characterized in settings where a predictive model would arguably have the most impact. Second, almost all of the AI models are, to some extent, trained on lung screening scans from the National Lung Screening Trial (NLST) (21), which raises the question of whether they generalize across institutions and to patients with incidentally detected nodules and metastases to the lung from other sites. Third, these models predict different outcomes. Some assess the risk of developing lung cancer over a multiyear period, whereas others estimate the probability that an observed pulmonary nodule is malignant. A comparative analysis of these models using a standardized outcome (eg, 2-year diagnosis of lung cancer) can inform "off-label" use of models but has not yet been performed. Finally, there is an urgent need to risk stratify intermediate-risk nodules and reduce the number of biopsies on nodules that appear indeterminate but are diagnosed as benign. To our knowledge, a systematic analysis of how existing models perform in this setting has not been performed.

This study aimed to evaluate eight validated lung cancer prediction models on cohorts with screening-detected nodules, incidentally detected nodules collected in both retrospective and prospective fashion, and nodules that underwent a bronchoscopic biopsy. These different settings are where we envision a well-designed predictive model will have a tangible impact on patient care. We implemented each model from their published code repositories and curated each cohort from their primary available source.

Materials and Methods

Cohorts

The data included in this retrospective study were sourced from the NLST, through the Cancer Data Access System, and our Vanderbilt University Medical Center (VUMC). This study also included data from the Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions (MCL), which includes the Veterans Affairs facility associated with VUMC, University of Pittsburgh Medical Center (UPMC), Detection of Early Lung Cancer Among Military Personnel (DECAMP), and University of Colorado Denver (UCD). We derived 10 named cohorts from these sites using different inclusion criteria (Table 1). We obtained CT scans, demographics, and questionnaire data from the CT arm of the NLST upon request with a data use agreement *(https://cdas.cancer.gov/learn/nlst/images/)*. Data from the Vanderbilt Lung Screening Program (VLSP) cohort were acquired under our home institutional review board supervision (no. 181279). Data from the Longitudinal Incidental VUMC (LI-VUMC) and BRONCH (biopsied nodules) cohorts were acquired under institutional review board supervision (no. 140274). Data from the VUMC cohort were acquired under institutional review board supervision (no. 030763 and 000616). Data from the UPMC, UCD, and DECAMP cohorts were acquired via academic collaborations under different grants. Acquisition of data from these cohorts was compliant with the Health Insurance Portability and Accountability Act. Regarding patients who have been previously reported, the NLST is a widely studied public dataset. Portions of the VLSP cohort have been reported in Li et al (22) (n = 1189) and Gao et al (20) (n = 147), and the LI-VUMC cohort was previously reported in Li et al (23) and Li et al (24). Patients from VUMC, PUMC, UCD, and DECAMP have been previously studied in Gao et al (19) (n = 1331) and Kammer et al (25) (n = 457).

Image Preprocessing

We used a documented pipeline (26) that includes algorithmic analysis and manual visual assessment to ensure every scan used in this study passed certain image quality standards (Fig 1). Specifically, we excluded scans with severe imaging artifacts, scans with a section thickness greater than or equal to 5 mm, and scans without the full lung field in the field of view. A total of 713 studies were excluded because of insufficient quality (Fig S1). Patient health information was removed using the MIRC Anonymizer (27).

Predictive Models

We selected an array of models for lung cancer prediction (Table 2), including models designed to estimate lung cancer risk (ie, Sybil) as well as models designed to predict the malignancy probability of a pulmonary nodule. We included the Brock (9) and Mayo (8) models because they are among the most cited, validated, and used in clinical practice. We studied several AI models incorporating a range of approaches that would allow us to examine the efficacy of three strategies: Liao et al (14) and Sybil (15) as single-time-point chest CT approaches, Distanced Long Short-

Table 1: Cohort Inclusion and Exclusion Criteria

Cohort	Inclusion and Exclusion Criteria
NLST-test	Patients correspond to the Ardila et al (18) test set. These patients were not evaluated by any of the predictive models in this study. Lung cancer events were the biopsy-confirmed lung cancers reported by the NLST. Patients without a confirmed outcome were excluded from this study.
NLST-test-nodule	Subset of the NLST-test cohort in which included patients had at least one positive nodule finding from their CT examinations as defined in the NLST (≥4 mm).
NLST-dev	All patients enrolled in the CT arm of the NLST and not part of NLST-test. Those without available imaging or without a confirmed outcome were excluded.
VLSP	Patients meeting the American Cancer Society criteria for lung screening and who were enrolled in the lung screening program at VUMC from 2015 to 2018. Patients receive longitudinal follow-up after a positive screen and lung cancer events were confirmed via biopsy reports. Nodule characteristics are missing because radiology reports were not available.
LI-VUMC	Patients from VUMC who underwent three chest CT examinations within 5 years between 2012 and 2019. These patients were identified through ICD codes to have a pulmonary nodule and no cancer before the nodule. We defined lung cancer outcomes through ICD codes representing any malignancy found in the bronchus or lung parenchyma, including metastases from other sites (23). Nodule characteristics are missing because radiology reports were not available.
MCL-VUMC	Prospectively enrolled patients from VUMC and its associated Veterans Administration facility between 2003 and 2017. Cohorts prefixed with "MCL-" meet the following inclusion criteria. Patients must be aged 18–80 years and were detected incidentally to have a pulmonary nodule with a diameter between 6 and 30 mm. Patients consented at initial nodule detection, and serum test and CT scan were acquired at that time. Longitudinal imaging and biopsy-confirmed diagnosis for malignant nodules were collected during a 2-year period following initial nodule detection.
MCL-UPMC	Prospectively enrolled patients from UPMC according to consortium inclusion criteria. Longitudinal imaging after initial nodule detection was not available.
MCL-DECAMP	Prospectively enrolled patients from 12 clinical centers associated with the DECAMP (36) study protocol. Of note, cases and controls are matched on nodule size.
MCL-UCD	Prospectively enrolled patients from UCD according to MCL inclusion criteria. Longitudinal imaging after initial nodule detection was not available.
BRONCH	Prospectively collected cohort of patients who underwent a bronchoscopic lung biopsy for a pulmonary nodule (defined as a lesion <3 cm) at the lung nodule clinic of VUMC between the years of 2017 and 2019. The subsequent biopsy report from the bronchoscopy was used to determine benign versus malignant status of the nodule.

Note.—BRONCH = cohort of patients who underwent bronchoscopic lung biopsy at VUMC, DECAMP = Detection of Early Lung Cancer Among Military Personnel, ICD = *International Classification of Diseases*, MCL = Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions, NLST = National Lung Screening Trial, UCD = University of Colorado Denver, UPMC = University of Pittsburgh Medical Center, VUMC = Vanderbilt University Medical Center.

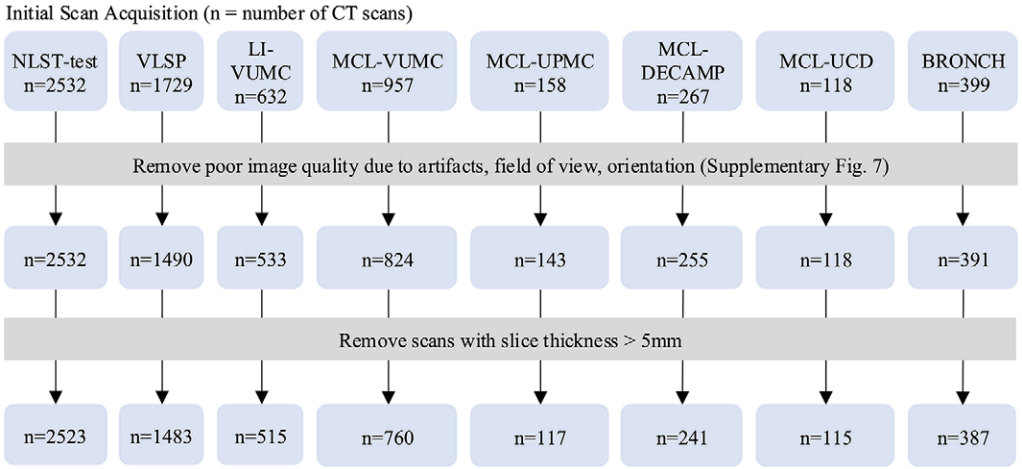

Figure 1: Flowchart of image quality assurance pipeline for each cohort. Exclusion criteria included severe artifact, nonstandard chest or body orientation, field of view that did not fully include the lung, and section thickness over 5 mm. BRONCH = cohort of patients who underwent bronchoscopic lung biopsy at VUMC, DECAMP = Detection of Early Lung Cancer Among Military Personnel, LI-VUMC = Longitudinal Incidental-VUMC, MCL = Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions, NLST = National Lung Screening Trial, UCD = University of Colorado Denver, UPMC = University of Pittsburgh Medical Center, VLSP = Vanderbilt Lung Screening Program, VUMC = Vanderbilt University Medical Center.

Table 2: Lung Cancer Predictive Model Characteristics

Model	Year Published	Input	Training Distribution	Cancer Prevalence	Outcome Criteria	Approach
Mayo (8)	1997	Age, PH, SS, NSpic, NUL, NSize*	Mayo Clinic (*n* = 419)	23%	2-year LC risk proven via tissue biopsy or no findings at follow-up	Logistic regression
Brock (9)	2013	Age, Sex, FH, Emp, Nsize, Nspic NUL, Ncount, Ntype†	PanCan (*n* = 1871)	5.5%	2-year LC risk proven via tissue biopsy or no findings at follow-up	Logistic regression
Liao et al (14)	2017	Single chest CT	NLST-dev (*n* = 5436)	17%	1-year LC risk proven via tissue biopsy or no findings at follow-up	ResNet, nodule detection, and ROI-based prediction
Sybil (15)	2023	Single chest CT	NLST-dev (*n* = 12 672)	17%	Up to 6-year LC risk proven via tissue biopsy or no findings at follow-up	ResNet, global chest features, and guided attention
DLSTM (16)	2020	Longitudinal chest CT	NLST-dev (*n* = 5436)	17%	6-year LC risk proven via tissue biopsy or no findings at follow-up	LSTM network, ROI-based prediction, encodes time interval between scans
TdViT (17)	2023	Longitudinal chest CT	NLST-dev (*n* = 5436)	17%	6-year LC risk proven via tissue biopsy or no findings at follow-up	Transformer network, ROI-based prediction, encodes time interval between scans
DeepLungScreening (20)	2021	Single chest CT, Age, Education, BMI, PH, FH, SS, Quit, PYR	NLST-dev (*n* = 5436)	17%	2-year LC risk proven via tissue biopsy or no findings at follow-up	ResNet, ROI-based prediction, late fusion of imaging and clinical features
DeepLungIPN (19)	2021	Single chest CT, Age, BMI, PH, SS, PYR, Nsize, NSpic, NUL, Serum biomarker‡	MCL cross-validation§ (*n* = 1232)	59%	2-year LC risk proven via tissue biopsy or no findings at follow-up	DeepLungScreening, serum biomarker

Note.—BMI = body mass index, CYFRA = cytokeratin 19 fragment, DECAMP = Detection of Early Lung Cancer Among Military Personnel, DLSTM = Distanced LSTM, FH = family history of lung cancer, LC = lung cancer, NLST = National Lung Screening Trial, NSize = nodule size, NSpic = nodule spiculation present or absent, NUL = nodule in the upper lobes, Ntype = nodule type, Ncount = number of nodules, PYR = pack-years of smoking, PH = personal history of any cancer, Emp = presence of emphysema, Quit = years since the person quit smoking, ROI = region of interest, SD = smoking duration, SI = smoking intensity (average number of cigarettes smoked a day), SS = smoking status (former vs current smoker), TdViT = Time-distance Vision Transformer.
* Largest nodule diameter (mm).
† Categorized as nonsolid or with ground-glass opacity, part-solid, and solid.
‡ Serum concentration of hs-CYFRA 21–1 (natural log of ng/mL) (37).
§ Combination of MCL-VUMC, MCL-DECAMP, MCL-UCD.

Term Memory (16) and Time-distance Vision Transformer (17) as a longitudinal chest CT approaches, and DeepLungScreening (20) and DeepLungIPN (19) as models with multimodal inputs.

We split patients with confirmed follow-up in the NLST into development (NLST-dev) and test (NLST-test) sets. NLST-test contains the patients in the Ardila et al (18) test set who had confirmed follow-up, and these scans remained unseen until evaluation. NLST-dev was used to retrain, from random weights, several of the models using a standardized 2-year lung cancer outcome. The purpose of retraining was to ensure that the models were not trained on NLST-test and to standardize the predicted outcome across each model. Specifically, the years between initial observation of the patient and the outcome, or year-to-outcome, was not standardized across the evaluated models (Table 2). Models developed using a shorter year-to-outcome have an easier task than models developed using a longer year-to-outcome. In this way, differences in year-to-outcomes can confound model comparisons. For longitudinal imaging models, the outcome was whether the patient was diagnosed with lung cancer within 2 years of the patient's latest scan. The logistic regression models were not retrained and were evaluated as published because they were already blinded to NLST-test and we did not have their original development dataset, which is needed to control for year-to-outcome. Sybil was also evaluated as published because the model was already blinded to NLST-test and its prediction includes a 2-year outcome. Last, DeepLungIPN was originally trained using a cross-validation of MCL-VUMC,

Table 3: Cohort Characteristics

Cohort	NLST-dev	NLST-test	NLST-test-nodule	VLSP	LI-VUMC	VUMC	UPMC	DECAMP	UCD	BRONCH
Program type	Screening	Screening	Screening	Screening	Screening, incidental	Incidental	Incidental	Incidental	Incidental	Bronchoscopy
Institution	MCL	MCL	MCL	VUMC	VUMC	VUMC, VA VUMC	UPMC	MCL	UCD	VUMC
Program period	2002–2009	2002–2009	2002–2009	2015–2018	2012–2021	2003–2017	2006–2015	2013–2017	2010–2018	2017–2019
No. of patients	5436	898	896	882	219	364	117	131	115	373
No. of patients with lung cancer	901 (17)	149 (17)	147 (16)	24 (3.0)	37 (17)	238 (65)	48 (41)	64 (49)	57 (50)	230 (62)
No. of scans	14748	2523	2440	1483	515	760	117	241	115	387
No. of scans with lung cancer	1866 (13)	313 (12)	298 (12)	51 (3.4)	50 (10)	517 (68)	48 (41)	100 (41)	57 (50)	240 (62)
Section thickness (mm)	2.1 ± 0.65	2.1 ± 0.42	2.1 ± 0.42	0.81 ± 0.21	0.77 ± 0.61	1.8 ± 1.1	2.2 ± 0.69	1.7 ± 0.91	1.4 ± 0.79	0.9 ± 0.38
Age (y)	62 ± 5.2	62 ± 5.2	62 ± 5.2	65 ± 5.8	59 ± 13	69 ± 11	68 ± 8.5	68 ± 7.9	66 ± 8.3	64 ± 12
Sex (male)	3270 (60)	546 (61)	546 (61)	483 (55)	109 (50)	165 (45)	49 (42)	30 (23)	31 (27)	168 (45)
BMI	28 ± 4.8	28 ± 4.9	28 ± 5.0	28.4 ± 6.0	27 ± 7.4	28 ± 6.5	28 ± 4.9	26 ± 5.4	29 ± 6.2	28 ± 6.8
Personal cancer history	256 (4.7)	43 (4.8)	43 (4.8)	135 (15)	NA	129 (35)	3 (2.6)	65 (50)	13 (11)	194 (52)
Family lung cancer history	1194 (22)	179 (20)	177 (20)	149 (17)	NA	41 (11)	0	0	10 (8.7)	88 (24)
Smoking status										
Never	0	0	0	0	NA	33 (9)	0	11 (8.4)	22 (19)	90 (24)
Former	2781 (51)	469 (52)	468 (52)	357 (40)	NA	195 (54)	76 (65)	67 (51)	54 (47)	214 (57)
Current	2655 (49)	429 (48)	428 (48)	525 (60)	NA	121 (33)	41 (35)	53 (40)	39 (34)	69 (18)
Smoking pack-years	56 ± 25	59 ± 28	59 ± 28	48 ± 21	NA	47 ± 33	48 ± 23	50 ± 25	50 ± 33	29 ± 30
Nodule size (mm)	8.0 ± 6.2	7.9 ± 5.9	7.9 ± 5.9	NA	NA	19 ± 13	16 ± 8.9	15 ± 7.0	18 ± 15	2.2 ± 1.3
Nodule count	1.2 ± 1.3	1.3 ± 1.2	1.3 ± 1.2	NA	NA	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
Nodule attenuation										
Solid	7494 (51)	1351 (54)	1351 (55)	NA	NA	725 (95)	99 (85)	241 (100)	115 (100)	325 (84)
Part-solid	507 (3.4)	69 (2.7)	69 (2.8)	NA	NA	21 (2.8)	18 (15)	0	0	51 (13)
Nonsolid or GGO	1439 (10)	241 (9.6)	241 (9.9)	NA	NA	14 (1.8)	0	0	0	11 (2.8)
Nodule spiculation (present)	997 (6.7)	200 (7.9)	200 (8.2)	NA	NA	229 (30)	15 (13)	126 (52)	30 (26)	173 (45)
Nodule location										
Upper lobe	5554 (38)	996 (39)	996 (41)	NA	NA	447 (59)	63 (54)	143 (59)	71 (62)	203 (52)
Lower lobe	4391 (30)	777 (31)	777 (32)	NA	NA	313 (41)	54 (46)	98 (41)	44 (38)	184 (48)

Note.—Data values are presented as means ± SDs or numbers of patients with percentages in parentheses. BMI = body mass index (calculated by dividing weight in kilograms by height in meters squared), BRONCH = cohort of patients who underwent bronchoscopic lung biopsy at VUMC, DECAMP = Detection of Early Lung Cancer Among Military Personnel, GGO = ground-glass opacity, LI-VUMC = Longitudinal Incidental-VUMC, MCL = Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions, NA = not available, NLST = National Lung Screening Trial, NLST-dev = NLST development set, NLST-test = NLST test set, NLST-test-nodule = subset of NLST-test of patients with at least one positive nodule finding, UCD = University of Colorado Denver, UPMC = University of Pittsburgh Medical Center, VLSP = Vanderbilt Lung Screening Program, VUMC = Vanderbilt University Medical Center, VA = Veterans Affairs.

MCL-DECAMP, and MCL-UCD. This model was evaluated as published because it includes a blood biomarker that was only collected in the MCL cohorts.

Implementation and training of models followed their original methodology unless otherwise specified. Details about the development site and training distribution are reported in Table

Table 4: Model Classification of *n*-year Lung Cancer Risk across Selected Cohorts

Classification		NLST-test (*n* = 898)	NLST-test-nodules (*n* = 896)	VLSP (*n* = 882)	LI-VUMC (*n* = 219)	MCL-VUMC (*n* = 364)	MCL-UPMC (*n* = 117)	MCL-DECAMP (*n* = 131)	MCL-UCD (*n* = 115)	BRONCH (*n* = 373)	Average Rank
Input	Model										
Clinical variables	Mayo	NA*	0.804 [0.798, 0.809]	NA*	NA*	0.706 [0.704, 0.708]	0.864 [0.862, 0.867]	0.568 [0.565, 0.571]	0.716 [0.712, 0.719]	0.621 [0.615, 0.628]	3.5 (7–1) *n* = 6
Clinical variables	Brock	NA*	0.789 [0.782, 0.796]	NA*	NA*	0.716 [0.714, 0.718]	0.885 [0.883, 0.886]	0.662 [0.659, 0.666]	0.713 [0.710, 0.716]	0.497 [0.494, 0.499]	3.2 (5–2) *n* = 6
Single CT AI	Liao et al	0.751 [0.747, 0.756]	0.755 [0.750, 0.759]	0.723 [0.712, 0.734]	0.644 [0.635, 0.653]	0.662 [0.660, 0.664]	0.779 [0.776, 0.782]	0.706 [0.703, 0.709]	0.660 [0.656, 0.663]	0.621 [0.614, 0.628]	3.9 (6–1) *n* = 9
Single CT AI	Sybil	0.881 [0.877, 0.885]†	0.879 [0.872, 0.885]†	0.779 [0.768, 0.789]	0.763 [0.756, 0.770]	0.700 [0.694, 0.706]	0.889 [0.884, 0.895]	0.606 [0.597, 0.616]	0.764 [0.756, 0.772]	0.623 [0.618, 0.629]	2.6 (6–1) *n* = 9
Longitudinal CT AI	DLSTM	0.738 [0.734, 0.743]	0.727 [0.721, 0.731]	NA‡	0.711 [0.702, 0.720]	0.743 [0.741, 0.745]	NA§	0.778 [0.774, 0.781]	NA§	NA§	3.6 (6–2) *n* = 5
Longitudinal CT AI	TDViT	0.797 [0.793, 0.802]	0.790 [0.785, 0.794]	NA‡	0.773 [0.764, 0.781]†	0.753 [0.750, 0.755]	NA§	0.823 [0.820, 0.825]†	NA§	NA§	1.8 (3–1) *n* = 5
Multimodal	DLS	0.783 [0.778, 0.788]	0.776 [0.771, 0.782]	0.810 [0.799, 0.820]†	NA‖	NA‖	NA‖	NA‖	NA‖	NA‖	2.7 (4–1) *n* = 3
Multimodal	DLI	NA‖	NA‖	NA‖	NA‖	0.856 [0.854, 0.858]†	0.936 [0.935, 0.938]†	0.742 [0.739, 0.745]	0.851 [0.849, 0.854]†	NA†	1.5 (3–1) *n* = 4

Note.—Except where indicated, data are bootstrapped mean areas under the receiver operating characteristic curve, with 95% CIs in brackets. The data in Average Rank are average, range, and the number of cohort evaluations performed. The *n*-year lung cancer risk for each cohort was 2-year risk for each cohort except LI-VUMC, which was 3-year risk, and BRONCH, which was 1-year risk. AI = artificial intelligence, BRONCH = cohort of patients who underwent bronchoscopic lung biopsy at VUMC, COPD = chronic obstructive pulmonary disease, CYFRA = cytokeratin 19 fragment, DLI = DeepLungIPN, DLS = DeepLungScreening, DECAMP = Detection of Early Lung Cancer Among Military Personnel, DLSTM = Distanced LSTM, LI-VUMC = Longitudinal Incidental-VUMC, MCL = Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions, NA = not available, NLST = National Lung Screening Trial, NLST-test = NLST test set, NLST-test-nodule = subset of NLST-test of patients with at least one positive nodule finding, TdViT = Time-distance Vision Transformer, UCD = University of Colorado Denver, UPMC = University of Pittsburgh Medical Center, VLSP = Vanderbilt Lung Screening Program, VUMC = Vanderbilt University Medical Center.

* Nodule characteristics unavailable (missing >10% of nodule size, attenuation, count, spiculation, or lobe location).

† Result was significantly different compared with each other method in the column for $P < .01$.

‡ Prohibitive class imbalance (only six of 23 patients with lung cancer have more than one scan).

§ No longitudinal imaging.

‖ Missing demographic, smoking history, COPD, or CYFRA covariates.

S1. Apart from removing scans that did not meet our image quality standards, we did not add or remove any pre- or postprocessing steps included in the models' pipeline. The code supporting model training, evaluation, and statistical analysis is available at *https://github.com/MASILab/lcancer_baselines*. Code supporting this project is made available at *https://github.com/MASILab/lcancer_baselines*.

Evaluation and Statistical Analysis

Evaluation included all the named cohorts except NLST-dev. A model-cohort evaluation was not feasible when a substantial portion of the input data was missing, specifically when an input variable was missing in more than 10% of cohort patients (Table S2). When an input variable was missing in less than 10% of patients, we conducted an evaluation using imputation based on a multivariable regression of the other available variables. The effect of imputation on model performance is shown in Table S3. We did not evaluate longitudinal imaging models (Distanced Long Short-Term Memory and Time-distance Vision Transformer) on cohorts in which longitudinal imaging was unavailable. When evaluating DeepLungIPN on the consortium cohorts, we report the out-of-fold cross-validation results.

We used AUC to measure model performance for classifying lung cancer cases and benign controls. For each model-cohort evaluation, we used a bootstrapping procedure to estimate the model's performance on the cohort's true population. The procedure drew 1000 samples of the same size with replacement from the original cohort. Each model's AUC was calculated for each sample, and we reported the mean AUC and 95% CI over all

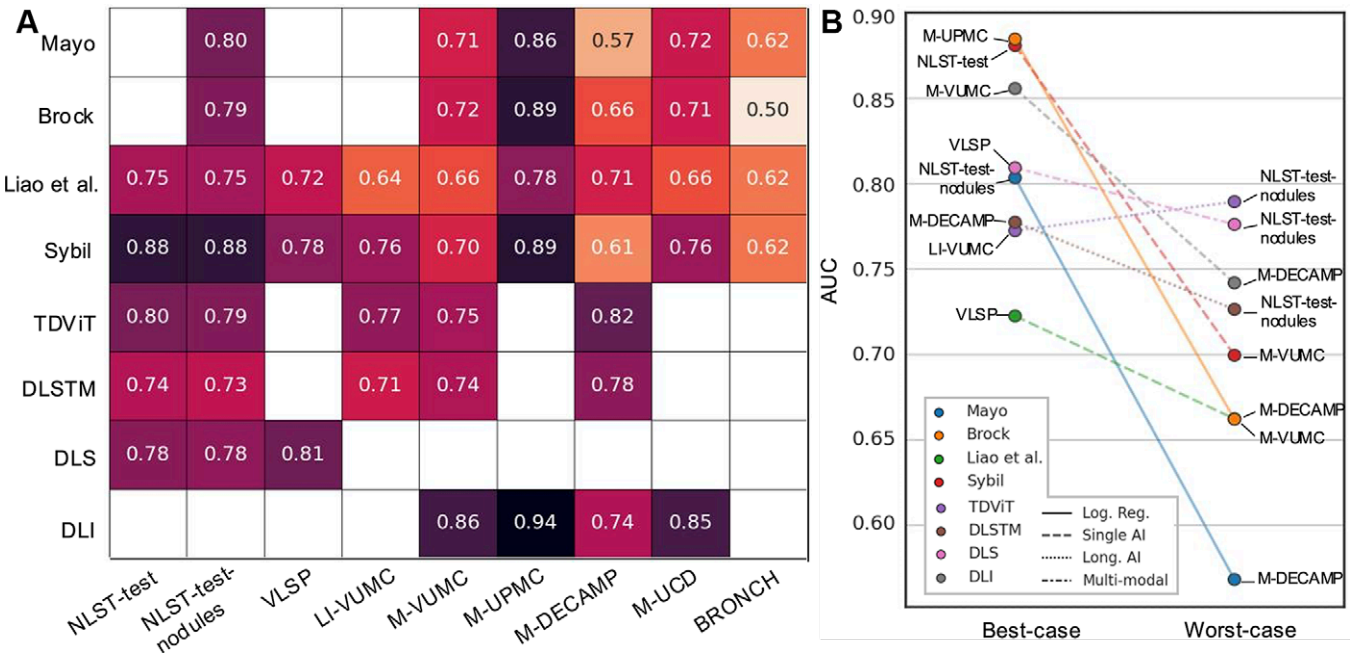

Figure 2: **(A)** Mean area under the receiver operating characteristic curve (AUC) for all lung cancer prediction models applied on all study cohorts. Almost all methods demonstrate a high degree of variance in performance across cohorts within most methods, which demonstrates the importance of contextualizing a model's performance by comparing it with the performance of baseline models. Darker shading indicates better AUC performance. The 95% CIs are shown on Figure S6. **(B)** Best- and worst-case performance for eight predictive models reveals robust performance of longitudinal and multimodal AI methods (ie, Time-distance Vision Transformer [TDViT], Distanced Long Short-Term Memory [DLSTM], DeepLungScreening [DLS], DeepLungIPN [DLI]) compared with other models. A model's worst-case performance is defined as its lowest ranked performance across all cohorts except BRONCH. BRONCH = cohort of patients who underwent bronchoscopic lung biopsy at VUMC, DECAMP = Detection of Early Lung Cancer Among Military Personnel, LI-VUMC = Longitudinal Incidental-VUMC, MCL = Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions, NLST = National Lung Screening Trial, UCD = University of Colorado Denver, UPMC = University of Pittsburgh Medical Center, VLSP = Vanderbilt Lung Screening Program, VUMC = Vanderbilt University Medical Center.

bootstrapped samples. A two-sided Wilcoxon signed rank test was used to evaluate significance of differences (significant at $P < .05$) in mean AUC between models within a single cohort. We did not test statistical differences across cohorts because patients were not paired across cohorts.

Model performance for non–small cell lung cancer (NSCLC) versus small cell lung cancer (SCLC) cases were compared in NLST-test nodules and MCL-VUMC. The mean AUC and 95% CI of each model was computed using the same bootstrap procedure drawn from the pool of patients with benign nodules and patients with either SCLC or NSCLC. An unpaired *t* test was used to evaluate whether the discrimination of a model of malignant versus benign was significantly different (significant at $P < .05$) with these two lung cancer subtypes.

An analysis of calibration before and after confidence correction was conducted. In each model-cohort evaluation, 10-fold cross-validation was used to fit isotonic regressions on the training set of each fold. Calibration was then evaluated on the validation set of each fold (Figs S2–S5). All statistical analyses were performed in Python version 3.8 with support from the SciPy package.

Results

Cohort Characteristics

The size of the evaluation cohorts ranged from 898 patients (NLST-test) to 115 patients (MCL-UCD). Mean age ranged from 59 years with an SD of 13 (LI-VUMC) to 69 years with an SD of 11 (MCL-VUMC). Male sex proportion ranged from as low as 23% (MCL-DECAMP) to as high as 61% (NLST-test). Cancer prevalence in the biopsied nodules cohort (BRONCH: 62%) was the highest, followed by incidentally detected nodules (MCL-VUMC: 65%, MCL-UPMC: 41%, MCL-DECAMP: 49%, MCL-UCD: 50%), and the screening cohorts (NLST-test: 17%, VLSP: 3%, LI-VUMC: 17%). Mean smoking pack-years fell within 47 to 59 years except for the biopsied nodules cohort, which had a mean pack-year of 29 years due to the high proportion of never-smokers. Overall cohorts were distributed differently in terms of cancer prevalence, demographics, smoking background, and nodule characteristics (Table 3).

Model Performance

Table 4 reports the mean AUCs and 95% CIs for each feasible model-cohort evaluation. Table S4 also reports corresponding sensitivity and specificity using an optimal cut-point for each model-cohort evaluation. Comparing the results rowwise reveals that almost all predictive models exhibited noticeable differences in performance across cohorts (Fig 2A).

The performance gaps were the largest between cohorts from different sites and different clinical settings (ie, Brock on NLST-test-nodules: 0.789 [0.782, 0.796] vs Brock on MCL-DECAMP: 0.662 [0.659, 0.66]). The performance gap remained large between cohorts from different sites but the same clinical setting (ie, Sybil on NLST-test: 0.881 [0.877, 0.885] vs Sybil on VLSP: 0.779 [0.768, 0.789]). In contrast, the performance gap between different cohorts from the same site but different clinical setting was generally smaller (ie, Liao et al on LI-VUMC: 0.644 [0.635, 0.653] vs Liao et al on MCL-VUMC: 0.662 [0.660, 0.664]).

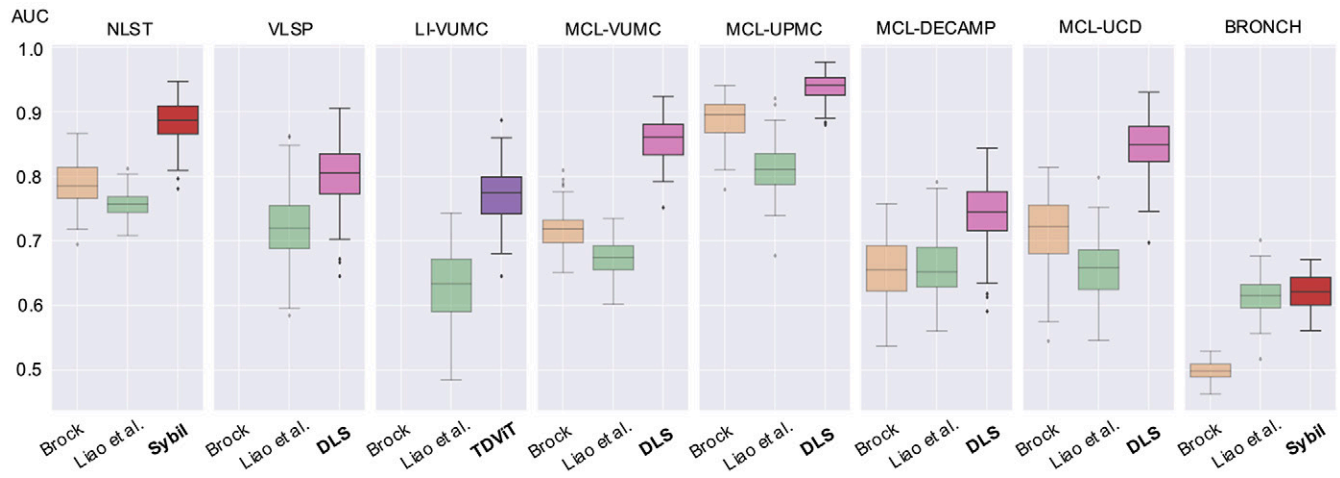

Figure 3: Box and whiskers plot shows areas under the receiver operating characteristic curves (AUCs) of 1000 bootstrapped samples from applying three selected methods across all study cohorts. Brock and Liao et al are selected as baselines to compare with the method achieving the highest classification performance in the corresponding cohort. The best performing method differs across cohorts. Among baselines and the best performers, bootstrapped AUC distributions demonstrate high variance across cohorts. DeepLungScreening (DLS) seems to perform the best in the cohorts with incidental nodules (MCL cohorts). Unsurprisingly, Time-distance Vision Transformer (TDViT) excels in the Longitudinal Incidental-VUMC imaging cohort (LI-VUMC). The box and line within the box denote the IQR and median, respectively. The whiskers denote 1.5 times the IQR, and points outside the whiskers denote outliers beyond this range. BRONCH = cohort of patients who underwent bronchoscopic lung biopsy at VUMC, DECAMP = Detection of Early Lung Cancer Among Military Personnel, MCL = Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions, NLST = National Lung Screening Trial, UCD = University of Colorado Denver, UPMC = University of Pittsburgh Medical Center, VLSP = Vanderbilt Lung Screening Program, VUMC = Vanderbilt University Medical Center.

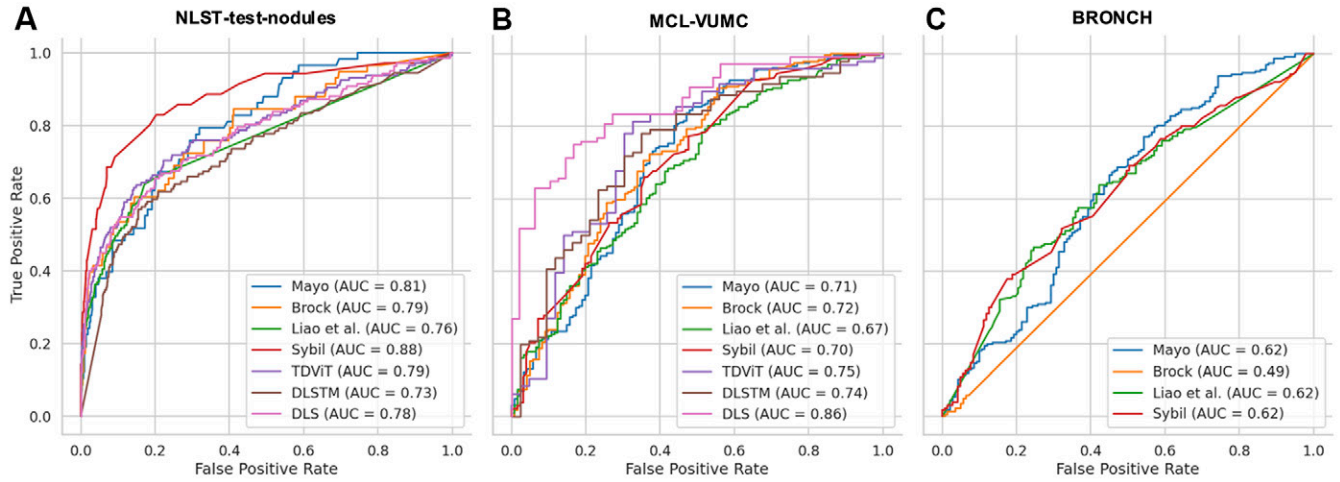

Figure 4: Receiver operating characteristic (ROC) curves demonstrate a failure to generalize across four select cohorts. Top performers in lung screening cohorts **(A)** are different than the top performers in cohorts with incidentally detected nodules **(B)**, and vice versa. **(C)** All evaluated models performed poorly on a retrospective cohort of patients selected to undergo diagnostic bronchoscopic biopsy (BRONCH) for a pulmonary nodule at VUMC. ROC curves for remaining evaluation cohorts are provided in Figure S7. DLS = DeepLungScreening, DLSTM = Distanced Long Short-Term Memory, MCL = Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions, TDViT = Time-distance Vision Transformer, VUMC = Vanderbilt University Medical Center.

Comparing the relative performances between multiple models across cohorts highlights the following findings. Single chest CT AI (Liao et al and Sybil) performed well in lung cancer screening cohorts (ie, Sybil model on NLST-test: AUC, 0.881 [95% CI: 0.887, 0.885]) (Fig 3A). These models were generally competitive with linear models while longitudinal and multimodal AI significantly outperformed linear models in every cohort. Results for MCL-UPMC represent this well, with Sybil (AUC, 0.889 [95% CI: 0.884, 0.895]) performing close to Brock (AUC, 0.885 [95% CI: 0.883, 0.886]), and DeepLungIPN (AUC, 0.936 [95% CI: 0.935, 0.938]) outstripping the performance of both.

Longitudinal or multimodal AI were top performers across all cohorts with incidental nodules (Figs 3, 4). They showed better worst-case performances in comparison to the other approaches (Fig 2B). Ranking the results within each cohort, we define a model's worst case as its lowest ranked performance across all cohorts except BRONCH. The worst-case performances of Distanced Long Short-Term Memory (on NLST-test nodules: AUC, 0.727 [95% CI: 0.721, 0.731]), Time-distance Vision Transformer (on NLST-test nodules: AUC, 0.790 [95% CI: 0.785, 0.794]), DeepLungScreening (on NLST-test nodules: AUC, 0.776 [95% CI: 0.7771, 0.782]), and DeepLungIPN (on MCL-DECAMP: AUC, 0.742 [95% CI: 0.739, 0.745]) were all moderate in terms of absolute AUC. In contrast, the worst-case performance of Mayo, Brock, Liao et al, and Sybil were low in terms of absolute AUC.

Models evaluated on the BRONCH cohort, representing nodules that are suspicious enough to warrant a biopsy, performed poorly, with mean AUCs ranging from 0.497 to 0.623 (Fig 4).

Table 5: Model Classification of Lung Cancer Risk by Cancer Subtype

	NLST-test Nodules*		MCL-VUMC†	
Model	SCLC (*n* = 18)	NSCLC (*n* = 119)	SCLC (*n* = 39)	NSCLC (*n* = 194)
Mayo	NA‡	0.810 (0.808, 0.812)	0.774 (0.772, 0.776)	0.683 (0.681, 0.685)
Brock	NA‡	0.792 (0.790, 0.794)	0.794 (0.792, 0.797)	0.688 (0.687, 0.690)
Liao et al	0.683 (0.680, 0.687)	0.770 (0.768, 0.771)	0.617 (0.614, 0.620)	0.688 (0.686, 0.690)
Sybil	0.728 (0.723, 0.733)§	0.899 (0.897, 0.900)§	0.701 (0.698, 0.703)	0.701 (0.699, 0.702)
DLSTM	0.663 (0.658, 0.668)	0.808 (0.806, 0.809)	0.730 (0.726, 0.735)	0.754 (0.751, 0.757)
TDViT	0.707 (0.702, 0.711)	0.771 (0.769, 0.773)	0.667 (0.661, 0.673)	0.760 (0.757, 0.763)
DLS	0.659 (0.654, 0.664)	0.792 (0.791, 0.794)	NA‖	NA‖
DLI	NA‖	NA‖	0.904 (0.901, 0.907)§	0.853 (0.851, 0.855)§

Note.—Data are bootstrapped mean areas under the receiver operating characteristic curve, with 95% CIs in parentheses. DLI = DeepLungIPN, DLS = DeepLungScreening, DLSTM = Distanced LSTM, NSCLC = non–small cell lung cancer, MCL = Consortium for Molecular and Cellular Characterization of Screen-Detected Lesions, NA = not available, NLST-test-nodule = subset of National Lung Screening Trial-test of patients with at least one positive nodule finding, SCLC = small cell lung cancer, TdViT = Time-distance Vision Transformer, VUMC = Vanderbilt University Medical Center.

* *n* malignant = 147, *n* benign = 749.

† *n* malignant = 238, *n* benign = 126

‡ Prohibitive class imbalance (*n* = 5).

§ Result was significantly different compared with every other method in the column for $P < .01$.

‖ Model evaluation not performed in these cases because the required covariates were missing (demographics, smoking history, chronic obstructive pulmonary disease, cytokeratin 19 fragment).

AI models discriminated NSCLC cases from benign findings better than SCLC cases (Table 5) in the lung screening setting (AUC for Sybil on NLST-test was 0.899 [95% CI: 0.897, 0.900] for NSCLC vs 0.728 [95% CI: 0.723, 0.733] for SCLC). In both lung screening and incidentally detected nodules, longitudinal models demonstrated better performance with NSCLC cases compared with SCLC cases (AUC for Time-distance Vision Transformer on MCL-VUMC was 0.760 [95% CI: 0.757, 0.763] for NSCLC vs 0.667 [95% CI: 0.661, 0.673] for SCLC).

Discussion

The most prominent result of our study was perhaps that there was no clear winner among the models evaluated. The performance of each model varied with site and clinical setting, which reflects a moderate degree of generalization failure that is often observed in both open-source and commercial predictive models across many medical domains (28). Those interested in using predictive models in lung cancer should be aware that these models, despite previous reports of successful external validation, most reliably achieve their expected performance when they are used in the same clinical context and site as they were developed in (29). Those involved in model deployment should consider fine-tuning models with a cohort that matches the site, clinical setting, and year-to-outcome in which the model will be used. Steps should be taken during model development to mitigate a failure to generalize when site and setting are unmatched with techniques such as image harmonization (30), fine-tuning (31), and potentially directly modeling the site-specific effects (30). These results motivate further investigation into the site- and context-specific factors that are driving a variance in performance and how they can be harmonized.

This study reveals the importance of interpreting a model's performance relative to the performance of other models on the same cohort. Doing so revealed several findings that were sustained across cohorts. Single chest CT AI (Liao et al and Sybil models) performed on par with linear models that included nodule variables (Mayo and Brock). As demonstrated previously (9,10,15), single chest CT AI is well suited for identifying individuals at risk for lung cancer who can benefit from starting or having more frequent lung imaging. Longitudinal and multimodal models demonstrated comparatively favorable performance on incidentally detected nodules. In contrast to other models, longitudinal and multimodal AI also appeared to be more robust across cohorts, as seen from their worst-case performances.

Given that nodules in the BRONCH cohort were inherently difficult to diagnose, the poor performance on this cohort was unsurprising. Because of missing data, we were not able to evaluate longitudinal and multimodal AI on BRONCH. A predictive model that is highly specific for lung cancer in this setting has the potential to prevent invasive management of benign nodules. Therefore, evaluation of longitudinal and multimodal AI on a retrospective cohort of biopsied nodules is a high priority area for future investigation.

Longitudinal imaging models performed better on NSCLC than SCLC. One explanation for this is that NSCLC is, on average, observed more frequently as an indeterminant nodule compared with faster progressing SCLC, which is often an advanced stage at first observation (32). These results warn that longitudinal imaging models may underperform on SCLC cases.

The regression calibrator improved calibration for most models evaluated on the NSLT, MCL, and BRONCH cohorts. Calibration remained poor or became worse for models evaluated

on highly imbalanced cohorts (VLSP and LI-VUMC), which align with previous findings (33).

Within the AI approaches, leveraging additional sources complementary data appears to be an effective strategy for improving classification performance. For instance, Sybil makes use of the entire chest, whereas Liao et al predicts cancer based on a few regions of interest, a technique that crops out portions of the lung field and discards the overall chest anatomy. Additionally, using longitudinal imaging that, when available, leads to performance gains across most of the cohorts. The integration of two or more consecutive chest CT studies allows the model to consider how imaging features change over time. The use of data from multiple modalities also appears to be effective. From a clinical perspective, the advantage of a multimodality approach is expected, because imaging findings are often interpreted in the context of the patient's clinical risk factors. The improved performance of longitudinal AI and multimodal AI in this study suggest that combining the two approaches is a promising direction.

We note the following limitations of our study. Because the evaluation cohorts are from 2002 to 2021, we expect different numerical results on cohorts drawn from current practice but a similar failure to generalize across clinical context and site. Several model-cohort evaluations were not conducted because of incomplete data. Extreme class imbalance in the model training cohort is another confounding factor that can affect a model's sensitivity and specificity. This is concerning for the Brock model, which was trained on a cohort with a cancer prevalence much smaller than those of other models. Other confounding sources include the differences in cohort size, scanner manufacturers, and scanner protocols (34,35). Finally, the evaluation of DeepLungIPN on its training cohort is limited because the results are from cross-validation. However, it still performed well when evaluated on a true external cohort (MCL-UPMC).

In summary, this study presents a comparative analysis of eight lung cancer prediction models against nine cohorts that represent clinically relevant use cases. Our results revealed a lack of generalized performance and showed that certain modeling strategies excelled in lung screening versus incidentally detected nodules, and all models fell short in a cohort with biopsied nodules. We highlight approaches in lung cancer predictive modeling that, if investigated further, have the potential to overcome these observed limitations.

Author affiliations:

[1] Medical Scientist Training Program, Vanderbilt University, Nashville, Tenn

[2] Department of Biomedical Engineering, Vanderbilt University, 2301 Vanderbilt Pl, Nashville, TN 37235

[3] Department of Computer Science, Vanderbilt University, Nashville, Tenn

[4] Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, Tenn

[5] Department of Digital Technology and Innovation, Siemens Healthineers, Princeton, NJ

[6] Division of Allergy, Pulmonary and Critical Care Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tenn

[7] Department of Thoracic Surgery, Vanderbilt University Medical Center, Nashville, Tenn

[8] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tenn

[9] Department of Radiology and Radiological Sciences, Vanderbilt University Medical Center, Nashville, Tenn

Received November 25, 2023; revision requested January 7, 2024; final revision received November 15; accepted January 15, 2025.

Address correspondence to: T.Z.L. (email: thomas.z.li@vanderbilt.edu).

Funding: This research was funded by the National Institutes of Health through grants F30CA275020, 2U01CA152662, and R01CA253923-02, the National Science Foundation Faculty Early Career Development grant 1452485 and National Science Foundation grant 2040462, the Vanderbilt Institute for Surgery and Engineering through grant T32EB021937-07, the Vanderbilt Institute for Clinical and Translational Research through grant UL1TR002243-06, and the Pierre Massion Directorship in Pulmonary Medicine.

Acknowledgments: We extend our gratitude to Heidi Chen, PhD, for their statistical guidance in service of this research.

Author contributions: Guarantors of integrity of entire study, **T.Z.L., A.K., Y.M., B.A.L.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **T.Z.L., K.X., M.N.K., F.M.**; clinical studies, **S.A., M.K.**; experimental studies, **T.Z.L., R.G., R.P., T.A.L.**; statistical analysis, **T.Z.L., M.N.K., T.A.L., B.A.L.**; and manuscript editing, **T.Z.L., K.X., A.K., R.G., M.N.K., D.X., Y.M., R.P., R.J.L., S.D., T.A.L., F.M., B.A.L.**

Data sharing: Data generated or analyzed during the study are available from the corresponding author by request.

Disclosures of conflicts of interest: T.Z.L. No relevant relationships. **K.X.** No relevant relationships. **A.K.** No relevant relationships. **R.G.** No relevant relationships. **M.N.K.** Grants or contracts, National Institutes of Health; chair of the Early Detection Research Network's Early Stage Investigator forum; member of the American Thoracic Society's Thoracic Oncology planning committee. **S.A.** No relevant relationships. **D.X.** No relevant relationships. **M.K.** No relevant relationships. **Y.M.** No relevant relationships. **R.P.** No relevant relationships. **R.J.L.** No relevant relationships. **S.D.** No relevant relationships. **E.L.G.** No relevant relationships. **T.A.L.** No relevant relationships. **K.L.S.** Grants or contracts, American Cancer Society, NIH/NCI; participation on Data Safety or Monitoring Board, RevealDx; stock or stock options, RevealDx. **F.M.** No relevant relationships. **B.A.L.** Funding, NIH, Vanderbilt University, Vanderbilt University Medical Center; consult to Silver Maple outside of the scope of this work as disclosed to and reviewed by Vanderbilt University; Editor-in-Chief of SPIE *Journal of Medical Imaging*; travel support related to SPIE role.

References

1. Rivera MP, Mehta AC, Wahidi MM. Establishing the diagnosis of 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):e142S–e165S.
2. Lokhandwala T, Bittoni MA, Dann RA, et al. Costs of Diagnostic Assessment for Lung Cancer: A Medicare Claims Analysis. Clin Lung Cancer 2017;18(1):e27–e34.
3. Massion PP, Walker RC. Indeterminate pulmonary nodules: risk for having or for developing lung cancer? Cancer Prev Res (Phila) 2014;7(12):1173–1178.
4. Gould MK, Tang T, Liu ILA, et al. Recent trends in the identification of incidental pulmonary nodules. Am J Respir Crit Care Med 2015;192(10):1208–1214.
5. Mazzone PJ, Lam L. Evaluating the Patient With a Pulmonary Nodule: A Review. JAMA 2022;327(3):264–273.
6. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: From the Fleischner Society 2017. Radiology 2017;284(1):228–243.
7. Detterbeck FC, Lewis SZ, Diekemper R, Addrizzo-Harris D, Alberts WM. Executive Summary: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 2013;143(5 Suppl):7S–37S.
8. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. Arch Intern Med 1997;157(8):849–855.
9. McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med 2013;369(10):910–919.
10. Tammemägi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. N Engl J Med 2013;368(8):728–736.
11. Revel MP, Bissery A, Bienvenu M, Aycard L, Lefort C, Frija G. Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? Radiology 2004;231(2):453–458.
12. Oxnard GR, Zhao B, Sima CS, et al. Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. J Clin Oncol 2011;29(23):3114–3119.
13. Devaraj A, van Ginneken B, Nair A, Baldwin D. Use of volumetry for lung nodule management: Theory and practice. Radiology 2017;284(3):630–644.

14. Liao F, Liang M, Li Z, Hu X, Song S. Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-OR Network. IEEE Trans Neural Netw Learn Syst 2019;30(11):3484–3495.
15. Mikhael PG, Wohlwend J, Yala A, et al. Sybil: A Validated Deep Learning Model to Predict Future Lung Cancer Risk From a Single Low-Dose Chest Computed Tomography. J Clin Oncol 2023;41(12):2191–2200.
16. Gao R, Tang Y, Xu K, et al. Time-distanced gates in long short-term memory networks. Med Image Anal 2020;65:101785.
17. Li TZ, Xu K, Gao R, et al. Time-distance vision transformers in lung cancer diagnosis from longitudinal computed tomography. Proc SPIE Int Soc Opt Eng 2023;12464:1246412.
18. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954–961. [Published correction appears in Nat Med 2019;25(8):1319.]
19. Gao R, Tang Y, Xu K, et al. Deep Multi-path Network Integrating Incomplete Biomarker and Chest CT Data for Evaluating Lung Cancer Risk. Proc SPIE Int Soc Opt Eng 2021;11596:115961E.
20. Gao R, Tang Y, Khan MS, et al. Cancer Risk Estimation Combining Lung Screening CT with Clinical Data Elements. Radiol Artif Intell 2021;3(6):e210032.
21. National Lung Screening Trial Research Team; Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365(5):395–409.
22. Li TZ, Hin Lee H, Xu K, et al. Quantifying emphysema in lung screening computed tomography with robust automated lobe segmentation. J Med Imaging (Bellingham) 2023;10(4):044002.
23. Li TZ, Xu K, Chada NC, et al. Curating Retrospective Multimodal and Longitudinal Data for Community Cohorts at Risk for Lung Cancer. medRxiv 2023.11.03.23298020 [preprint] 2023.11.03.23298020. https://www.medrxiv.org/content/10.1101/2023.11.03.23298020. Posted November 4, 2023. Accessed November 7, 2023.
24. Li TZ, Still JM, Xu K, et al. Longitudinal Multimodal Transformer Integrating Imaging and Latent Clinical Signatures From Routine EHRs for Pulmonary Nodule Classification. arXiv 2304.02836 [preprint] https://arxiv.org/abs/2304.02836. Posted April 6, 2023. Accessed May 20, 2023.
25. Kammer MN, Lakhani DA, Balar AB, et al. Integrated Biomarkers for the Management of Indeterminate Pulmonary Nodules. Am J Respir Crit Care Med 2021;204(11):1306–1316.
26. Gao R, Khan MS, Tang Y, et al. Technical Report: Quality Assessment Tool for Machine Learning with Clinical CT. https://www.vumc.org/radiology/lung. Accessed June 13, 2023.
27. The MIRC DICOM Anonymizer. MircWiki. https://mircwiki.rsna.org/index.php?title=The_MIRC_DICOM_Anonymizer. Accessed November 22, 2023.
28. Lasko TA, Strobl EV, Stead WW. Why Do Clinical Probabilistic Models Fail To Transport Between Sites? arXiv 2311.04787 [preprint] https://arxiv.org/abs/2311.04787. Posted November 8, 2023. Accessed November 8, 2023.
29. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. Nat Med 2023;29(11):2686–2687.
30. Hu F, Chen AA, Horng H, et al. Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. Neuroimage 2023;274:120125.
31. Li TW, Lee GC. Performance Analysis of Fine-tune Transferred Deep Learning. In: 2021 IEEE 3rd Eurasia Conference on IOT, Communication and Engineering (ECICE). IEEE, 2021; 315–319.
32. Rudin CM, Brambilla E, Faivre-Finn C, Sage J. Small-cell lung cancer. Nat Rev Dis Primers 2021;7(1):3.
33. Kull M, Silva Filho TM, Flach P. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. Electron J Stat 2017;11(2):5052–5080.
34. Li Y, Lu L, Xiao M, et al. CT Slice Thickness and Convolution Kernel Affect Performance of a Radiomic Model for Predicting EGFR Status in Non-Small Cell Lung Cancer: A Preliminary Study. Sci Rep 2018;8(1):17913.
35. Choe J, Lee SM, Do KH, et al. Deep Learning-based Image Conversion of CT Reconstruction Kernels Improves Radiomics Reproducibility for Pulmonary Nodules or Masses. Radiology 2019;292(2):365–373.
36. Billatos E, Duan F, Moses E, et al. Detection of early lung cancer among military personnel (DECAMP) consortium: study protocols. BMC Pulm Med 2019;19(1):59.
37. Kammer MN, Kussrow AK, Webster RL, et al. Compensated Interferometry Measures of CYFRA 21-1 Improve Diagnosis of Lung Cancer. ACS Comb Sci 2019;21(6):465–472.