Cancer Risk Estimation Combining Lung Screening CT with Clinical Data Elements

Riqiang Gao, ME • Yucheng Tang, MS • Mirza S. Khan, MD • Kaiwen Xu, MS • Alexis B. Paulson, MS • Shelbi Sullivan, BS • Yuankai Huo, PhD • Stephen Deppen, PhD • Pierre P. Massion, MD • Kim L. Sandler, MD • Bennett A. Landman, PhD

From the Departments of Computer Science (R.G., K.X., Y.H., B.A.L.) and Electrical and Computer Engineering (Y.T., Y.H., B.A.L.), Vanderbilt University, 400 24th Ave S, Featheringill Hall, Room 371, Nashville, TN 37235; and Departments of Radiology and Radiological Sciences (A.B.P., K.L.S.), Thoracic Surgery (S.S., S.D.), General Internal Medicine and Public Health (M.S.K.), Biomedical Informatics (M.S.K.), and Medicine, Division of Allergy, Pulmonary and Critical Care Medicine (P.P.M.), Vanderbilt University Medical Center, Nashville, Tenn. Received January 25, 2021; revision requested March 24; revision received September 20; accepted September 28. **Address correspondence to** R.G. (e-mail: *riqiang.gao@vanderbilt.edu*).

This research was supported by National Science Foundation CAREER 1452485 and R01 EB017230. P.P.M. supported in part by the National Cancer Institute (UO1 CA196405). This study was supported in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, Tenn. This project was supported in part by the National Center for Research Resources, grant UL1 RR024975-01, and is now at the National Center for Advancing Translational Sciences, grant 2 UL1 TR000445-06. The de-identified imaging dataset(s) used for the analysis described were obtained from ImageVU, a research resource supported by the Vanderbilt Institute for Clinical and Translational Research Clinical Translational Science Award (ULTR000445 from NCATS/NIH), Vanderbilt University Medical Center institutional funding and Patient-Centered Outcomes Research Institute (PCORI; contract CDRN-1306-04869). This study was funded in part by the Martineau Innovation Fund Grant through the Vanderbilt-Ingram Cancer Center Thoracic Working Group and NCI Early Detection Research Network 2U01CA152662 to P.P.M.

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

Radiology: Artificial Intelligence 2021; 3(6):e210032 • https://doi.org/10.1148/ryai.2021210032 • Content codes: AI CH CT

Purpose: To develop a model to estimate lung cancer risk using lung cancer screening CT and clinical data elements (CDEs) without manual reading efforts.

Materials and Methods: Two screening cohorts were retrospectively studied: the National Lung Screening Trial (NLST; participants enrolled between August 2002 and April 2004) and the Vanderbilt Lung Screening Program (VLSP; participants enrolled between 2015 and 2018). Fivefold cross-validation using the NLST dataset was used for initial development and assessment of the co-learning model using whole CT scans and CDEs. The VLSP dataset was used for external testing of the developed model. Area under the receiver operating characteristic curve (AUC) and area under the precision-recall curve were used to measure the performance of the model. The developed model was compared with published risk-prediction models that used only CDEs or imaging data alone. The Brock model was also included for comparison by imputing missing values for patients without a dominant pulmonary nodule.

Results: A total of 23 505 patients from the NLST (mean age, 62 years ± 5 [standard deviation]; 13 838 men, 9667 women) and 147 patients from the VLSP (mean age, 65 years ± 5; 82 men, 65 women) were included. Using cross-validation on the NLST dataset, the AUC of the proposed co-learning model (AUC, 0.88) was higher than the published models predicted with CDEs only (AUC, 0.69; $P < .05$) and with images only (AUC, 0.86; $P < .05$). Additionally, using the external VLSP test dataset, the co-learning model had a higher performance than each of the published individual models (AUC, 0.91 [co-learning] vs 0.59 [CDE-only] and 0.88 [image-only]; $P < .05$ for both comparisons).

Conclusion: The proposed co-learning predictive model combining chest CT images and CDEs had a higher performance for lung cancer risk prediction than models that contained only CDE or only image data; the proposed model also had a higher performance than the Brock model.

Supplemental material is available for this article.

Lung cancer is one of the most prevalent cancers in the world and accounts for the highest cancer-related mortality in the United States (1) and worldwide. The findings of the National Lung Screening Trial (NLST) (2) led the United States Preventive Services Task Force to recommend annual lung cancer screening with low-dose CT for individuals between 50 and 80 years of age who *(a)* have over 30 pack-years of smoking and *(b)* quit tobacco use within the past 15 years or are current tobacco users (3). Separately, Tammemägi et al (4) proposed a refined risk stratification model integrating 11 clinical data elements (CDEs). Together, these guidelines and models can be used to estimate lung cancer risk before CT screening.

Manually human-curated imaging semantic features (eg, nodule size) and CDEs have been integrated for cancer risk estimation (5–7). The Brock University Pan-Canadian Early Detection of Lung Cancer Study (or PanCan) study (ie, Brock model) (7) is a logistic regression model incorporating imaging semantic features and CDEs for cancer risk estimation after nodule discovery. However, the acquisition and management of imaging semantic features require manual efforts from radiologists. In addition, the majority of individuals participating in lung cancer screening programs do not have lung nodules large enough to be documented. The potential for missing information makes it difficult to compute the lung cancer risk for these patients

Abbreviations

AUC = area under the receiver operating characteristic curve, AUPRC = area under the precision-recall curve, CDE = clinical data element, LIDC-IDRI = Lung Image Database Consortium and Image Database Resource Initiative, Lung-RADS = Lung Reporting and Data System, NLST = National Lung Screening Trial, PLCO = Prostate, Lung, Colorectal, and Ovarian, 3D = three dimensional, VLSP = Vanderbilt Lung Screening Program

Summary

A machine learning model integrating whole CT images and clinical data elements was developed that had better performance in determining lung cancer risk than models using risk factors or images only.

Key Points

- The proposed co-learning model (using both CT images and risk factors) had a higher performance (area under the curve [AUC], 0.88) for predicting lung cancer risk compared with existing representative models trained on CT images alone (AUC, 0.86; $P <$ .05) or on risk factors alone (AUC, 0.69; $P <$ .05).
- The co-learning model also had a higher AUC (0.91) than an established nodule risk calculator (Brock model [AUC range, 0.78–0.80]).

Keywords

Computer-aided Diagnosis (CAD), CT, Lung, Thorax

using methods that rely on imaging semantic features (eg, Brock model [7]), and these methods may therefore miss early-stage cancer. Several lung cancer CT screening studies, including the NLST and Dutch-Belgian Randomized Lung Cancer Screening (or NELSON) trial, were used to derive positivity criteria for the Lung Reporting and Data System (Lung-RADS) (8–10). In the NLST population, the cancer rate in patients with Lung-RADS 1 and Lung-RADS 2 is less than 1%, but the cancer rate is greater than 1% for patients with Lung-RADS 3 and greater than 10% for patients with Lung-RADS 4.

Deep learning techniques are transforming the medical imaging field (11) as the result of the success in general computer vision (12–15). Some works apply convolutional neural networks to either pulmonary nodule detection (16) or classification of benign versus malignant nodules (17–19). Recently, deep learning methods (20–22) have been developed for estimating cancer risk based on whole-chest CT scans. For example, Liao et al (20) won the Kaggle challenge (23) with a three-dimensional (3D) deep neural network consisting of two modules: suspicious nodule detection and classification.

We hypothesized that CT and CDEs provide complementary information for lung cancer risk estimation. In this study, we developed a model to integrate CT image features and CDEs in a unified machine learning framework. Specifically, we adopted deep learning techniques to extract quantitative imaging features and to train a co-learning deep learning model end-to-end by inputting CT imaging features and CDEs. The preprocessing and feature extraction of CT images were adapted from the work of Liao et al (20) and the CDE selection by the Prostate, Lung, Colorectal, and Ovarian (PLCO) $\text{PLCO}_{\text{M2012}}$ model (4). We evaluated our method using cross-validation on the NLST dataset and performed external testing using data from the Vanderbilt Lung Screening Program (VLSP) (*https://www.vumc.org/radiology/lung*).

Materials and Methods

Patient Selection

The in-house program with existing de-identified data was performed in accordance with the Health Insurance Portability and Accountability Act with approval from our institutional review board (181279). Two screening cohorts were used in this study. The NLST dataset is a large-scale randomized controlled trial for early diagnosis of lung cancer conducted in the United States, with approximately 54 000 participants enrolled between August 2002 and April 2004 (2). The VLSP is a comprehensive program that offers annual lung screening CT and management by the department of radiology at the Vanderbilt University Medical Center. Patients from VLSP included in this current study were enrolled between 2015 and 2018, and the requirement for written informed consent was waived. A portion of the patient population was studied previously (22,24,25), but those studies used only CT images for prediction. Patient demographics from NLST and VLSP included in this study are shown in Table 1. Screening eligibility criteria of the NLST and VLSP are comparable (Table 2).

Motivated by the $\text{PLCO}_{\text{M2012}}$ model (4), we included in our model as CDEs the following data elements defined in $\text{PLCO}_{\text{M2012}}$: age, education, body mass index, personal cancer history, family lung cancer history, tobacco use status, tobacco use quit time, and pack-years.

In addition to limiting inclusion in our study to patients who meet the eligibility criteria defined in Table 2, we included samples (64 898 scans; Table 1) for which we were able to successfully obtain source data and in which imaging data met the following criteria: *(a)* meeting quality standards (*https://github.com/MASILab/QA_tool*) (26), *(b)* achieving successful preprocessing as described by Liao et al (20), and *(c)* meeting the criteria of defining a positive or negative case. A positive case was defined as a biopsy-confirmed diagnosis of lung cancer within 2 years of the imaging date. A negative case was defined as a biopsy that was not consistent with a lung cancer diagnosis or stable radiographic findings for 2 or more years.

In our in-house dataset, we excluded 740 patients (with the use of CT) for whom definitive confirmation of cancer status is lacking or for whom data are missing (many of these are expected to be negative cases). Thus, the cancer rate of the in-house cohort is different from that of general screening cohorts.

CT Acquisition

The NLST data information can be found in National Lung Screening Trial Research Team et al (2). For VLSP, the section thickness (0018, 0050) is 1.0, the kilovoltage (0018, 0060) is 120.0, data collection diameter (0018, 0090) is 500.0, the reconstruction diameter (0018, 1100) is 447.0, the acquisition type (0018, 9302): "SPIRAL", the manufacturer (0008, 0070) is "Philips", and the scan options (0018, 0022) is "HELIX".

Table 1: Demographics of NLST and VLSP Used in This Study (Scan-Level)

Screening Program	NLST	VLSP
No. of patients	23 505	147
No. of patients with cancer	722 (3.1)	21 (14.3)
No. of scans	64 898	220
No. of scans with cancer	1037 (1.6)	40 (18.2)
Age (y)	62 ± 5	65 ± 5
No. of men/women	13 838/9667	82/65
BMI (kg/m^2)	28.07 ± 5.01	28.28 ± 5.68
COPD	1196 (5.1)	41 (27.9)
Personal cancer history	972 (4.1)	30 (20.4)
Family lung cancer history	5103 (21.7)	38 (25.9)
Tobacco use status		
Former	12 321	60
Current	11 184	87
Pack-years	55.58 ± 23.12	48.94 ± 19.82
Tobacco use quit time (y)	4.66 ± 5.62	3.33 ± 6.22
Education		
Less than high school	6904 (29.4)	6 (4.1)
High school graduate or GED	3309 (14.1)	29 (19.7)
Post-high school training, excluding college	5450 (23.2)	5 (3.4)
Associate's degree	4002 (17.0)	37 (25.2)
Bachelor's degree	3386 (14.4)	35 (23.8)
Graduate	425 (1.8)	35 (23.8)
Race		
White	21 801 (92.8)	134 (91.2)
Black	1030 (4.4)	12 (8.2)
Asian	518 (2.2)	0
Pacific	82 (0.3)	0
Latino	0	1 (0.7)
Indian	74 (0.3)	0

Note.—Values shown as either number with percentage in parentheses or mean ± standard deviation. BMI = body mass index, COPD = chronic obstructive pulmonary disease, GED = generalized education development certificate, NLST = National Lung Screening Trial, VLSP = Vanderbilt Lung Screening Program.

Table 2: Inclusion and Exclusion Criteria in NLST and VLSP

Screening Program	NLST	VLSP
Age range (y)	55–74	55–80
Tobacco use status	Current or former tobacco use	Current or former tobacco use
Pack-years	≥ 30	≥ 30
Quit smoking time (y)	Quit tobacco use ≤ 15	Quit tobacco use ≤ 15
Specific exclusion criteria		
	Prior lung cancer	Cancer diagnosis within past 5 years and current surveillance with chest CT
	Chest CT within 18 months	Signs or symptoms of lung cancer and/or pneumonia at the time of screening
	Hemoptysis	
	Unexplained weight loss of ≥15 lb in prior year	

Note.—NLST = National Lung Screening Trial, VLSP = Vanderbilt Lung Screening Program.

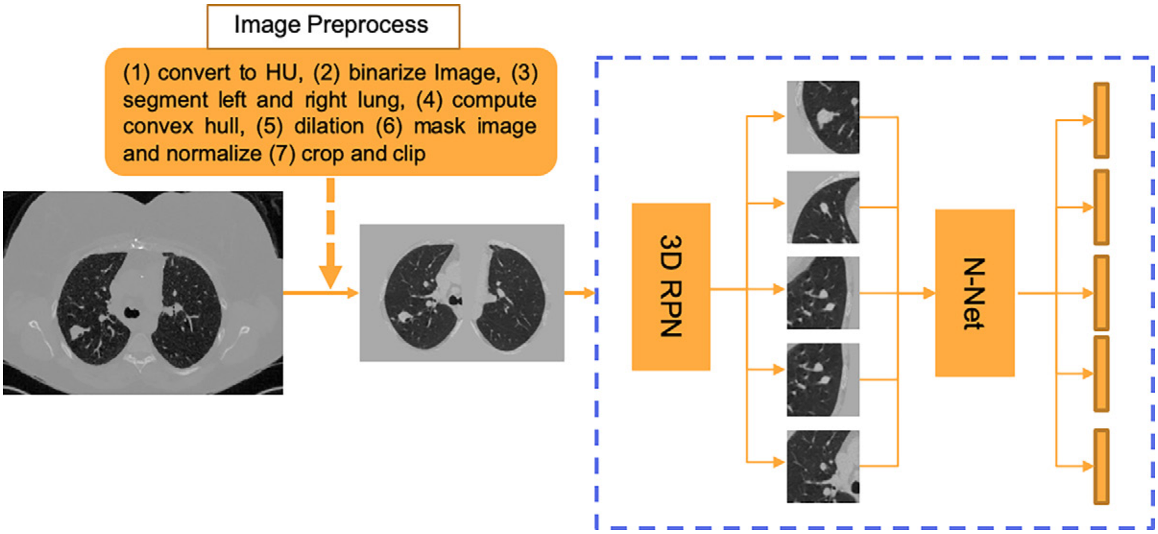
Image Preprocess
(1) convert to HU, (2) binarize Image, (3) segment left and right lung, (4) compute convex hull, (5) dilation (6) mask image and normalize (7) crop and clip
3D RPN
N-Net

Figure 1: Flowchart of image preprocessing and feature extraction. First, the raw CT image was preprocessed using the pipeline in Liao et al (20). Then, the top five confidence score regions were selected by a nodule detection network (three-dimensional region proposal network) pretrained by the Liao model. Each nodule region was extracted to a 1 × 128 dimension feature by a neural net (N-Net) pretrained by the Liao model. RPN = region proposal network.

Image Preprocessing and Pulmonary Nodule Detection

The CT image preprocessing and pulmonary nodule detection were adopted from Liao et al (20) with the publicly available code (*https://github.com/lfz/DSB2017*). Briefly, the image preprocessing uses the morphologic methods to segment the lung, clip the raw data matrix of CT images with intensity window [−1200, 600] Hounsfield units, and linearly transform data to [0, 255]. The pulmonary nodules were detected by a pretrained 3D region proposal network (12), and the nodules of the top five confidence scores were selected for the next steps. The confidence score was obtained from the nodule detection model, which reflected the likelihood of being a nodule of the selected region. Liao et al (20) validated that the top five regions were sufficient to capture most nodules. We follow the strategy described in Liao et al (20) by inputting the all-zeros-matrix to the deep learning network to extract the "nodule feature" if no nodule was detected. The illustration of image preprocessing and pulmonary nodule detection are shown in Figure 1.

Algorithm Design

The framework of our co-learning model was adapted from previous work by Gao et al (27), as shown in Figure 2. The input of the framework is a CT image and associated CDEs from a patient, and the output is a predicted lung cancer risk. As described in the previous section, the five proposals with top confidence score were obtained with a nodule detection model. Each nodule proposal was further converted to a 1 × 128 dimension feature vector. Patients were considered to have lung cancer if any nodule was malignant; therefore, the lung cancer prediction was formed as a multiple instance learning problem (20) wherein nodule proposals are instances. In a multiple instance learning model, the input is multiple instances from a sample and the output is the prediction of the sample. The sample is labeled as positive if any of the instances are positive; the sample is negative when all the instances are negative.

We use the attention-based multiple instance layer motivated by the approach developed by Ilse et al (28) to handle the multi-instance learning task (Fig 2). Briefly, the top five proposal features from one CT are converted to a single feature vector by the attention-based multiple instance layer, which uses a weighted average of instances (low-dimensional embeddings) whereby weights are determined by a neural network. The high-level features from CDEs were extracted by a deep learning subnetwork and then concatenated with the image feature for co-learning. After then using two fully connected layers, our model can predict the lung cancer risk in the range of (0, 1), where the risk levels of 0 and 1 represent cancer-free and cancer, respectively. Our model was trained with cross-entropy losses. *Cross-entropy* is a term from information theory, which builds on entropy and generally calculates the difference between two distributions. Cross-entropy loss is commonly used as a loss function to measure the difference of target and prediction in machine learning (ie, to measure how well they matched).

Model Comparisons

We compared our co-learning model with the popular CDE-only method $\text{PLCO}_{\text{M2012}}$ (4) and the image-only method of Liao et al (20). The $\text{PLCO}_{\text{M2012}}$ model uses 11 different epidemiologic CDEs for risk prediction. The Brock model was also included for comparison with appropriate imputation on those patients for whom imaging semantic features were absent. The Liao model (20) takes only the CT image as the input. We applied the pretrained model (20) (*https://github.com/lfz/DSB2017*) that won the Kaggle challenge (23) for comparison (ie, the Liao model). In NLST, the nodule an-

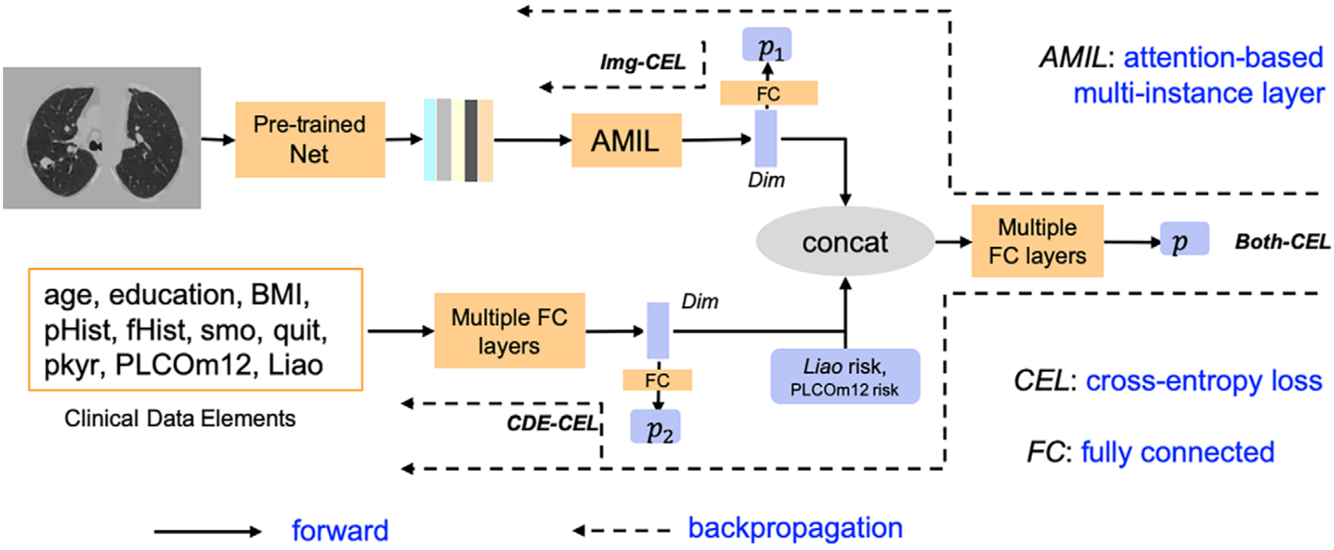

Figure 2: Our proposed co-learning framework. The "Pre-trained Net" is illustrated in Figure 1. Attention-based multi-instance layers were applied to combine the features from the top five nodule proposals. In the final step, clinical factors were concatenated (concat) with the image feature and followed with dense layers for the final prediction. BMI = body mass index, CDE = clinical data element, PLCO = Prostate, Lung, Colorectal, and Ovarian.

Table 3: AUC and AUPRC on the NLST Validation Dataset

Methods	AUC	AUPRC
No skill*	0.50	0.016
$\text{PLCO}_{\text{M2012}}$	0.69 ± 0.02	0.038 ± 0.006
Brock model	0.84 ± 0.01†	0.27 ± 0.03
Liao model	0.86 ± 0.02	0.32 ± 0.02
Our method	0.88 ± 0.02	0.34 ± 0.02

Note.—Values shown as either mean or mean ± standard deviation. AUPRC = area under the precision-recall curve, AUC = area under the receiver operating characteristic curve, NLST = National Lung Screening Trial, PLCO = Prostate, Lung, Colorectal, and Ovarian.

* No skill represents predicting without any knowledge, equivalent to random guessing.

† The AUC of the Brock model is computed by padding the default values (nodule size: 2 mm; speculation: no; upper lobe: no; nodule type: nonsolid) when factors are not available. Note that only nodule size smaller than 4 mm are missing in the NLST dataset.

notations are missing when the nodule size is less than 4 mm. The VLSP dataset did not contain nodule annotations when the lung scan was categorized as Lung-RADS 1 or 2.

Data Imputation for Brock Model

We imputed the missing values according to the criteria of Lung-RADS (30) and clinical experiences. We imputed the scan in Lung-RADS 1 as {nodule size: 0 mm; spiculated: no; nodule type: solid} and in Lung-RADS 2 as {nodule size: 3 mm; spiculated: no; nodule type: solid}. The missed "upper lobe" variable was achieved by a logistic regression trained with the available ones in NLST. To test if the model prediction was sensitive to imputed values, we included different imputation combinations for the Brock model. The nodule count was set as 1 when computing the Brock model. Note that the imputed values were only applied when computing the Brock model, as other compared methods (including ours) did not need human-curated radiomic features.

Cross-Validation on NLST

We randomly split the NLST cohort (Table 1) into five folds. Fivefold cross-validation (see Appendix E1 [supplement]) was applied in the NLST dataset, as shown in Table 3. In each fold of the evaluation, 20% of the cohort (ie, one fold) was held out from training as the test set, and the remainder of the data were split as 3:1 for training and tuning. The splitting was random. The model selection was based on the tuning set.

External Testing on the VLSP

To test the generalizability of our model, external validation was performed (ie, the model was trained with NLST and tested with VLSP). In external testing, we applied the same models developed during the cross-validation training on NLST without fine-tuning on VLSP. The prediction on VLSP was based on the average of the predictions of the fivefold models.

In the VLSP dataset, nodule annotations were not reported for nodules smaller than 6 mm in diameter. We compared performance when imputing other nodule diameter values (Table 4). We included four more combinations of imputed nodule size values for Lung-RADS 1 and Lung-RADS 2, and the imputed value combinations were based on the definition of Lung-RADS. As with the NLST, the nodule count was set as 1 when computing the Brock model. The nodule size, spiculation, and location of the primary nodule (upper lobe) were imputed with the latest CT records of a patient if values from the current CT were missing and values from the last CT are available, which may happen because some patients initially classified as having Lung-RADS 3 or 4 can be reclassified to Lung-RADS 2 if the nodule was stable. The receiver operating characteristic and precision-recall characteristic curves and their area under the receiver operating characteristic curve (AUC) values were computed with averaging prediction of five models from different folds.

Statistical Analysis

The predicted performance was evaluated based on the AUC and the area under the precision-recall curve (AUPRC) and the corresponding 95% CIs.

We show both bootstrapped 95% CIs and *P* value for indicating a statistical difference. The bootstrapped two-tailed test and the DeLong test (29) were used to compare the performance between the different models. No different conclusions were observed between those two tests (ie, reported $P < .05$ values were verified with both the bootstrapped two-tailed test and the DeLong test). The computation of the bootstrapped two-tailed test and 95% CIs were adapted from *https://github.com/mateuszbuda/ml-stat-util*.

Table 4: Comparison of Imputed Values for the Developed Model Compared with the Brock Model on the External Testing VLSP Dataset

Imputed Nodule Size (mm) [Lung-RADS 1, Lung-RADS 2]	Brock Model (7)	Our Method
[0, 2]	0.80 (0.71, 0.89)	0.91 (0.85, 0.95)
[0, 3]	0.78 (0.68, 0.88)	0.91 (0.85, 0.95)
[1, 3]	0.78 (0.68, 0.88)	0.91 (0.85, 0.95)
[1, 4]	0.78 (0.67, 0.88)	0.91 (0.85, 0.95)

Note.—Values are shown as mean with 95% CI in parentheses. Lung-RADS = Lung Reporting and Data System, VLSP = Vanderbilt Lung Screening Program.

Model Availability

We released our system at *https://github.com/AnonHappySubmit/DeepLungSceening_blind.* This link includes the source code, usage tutorial, pretrained deep learning models, docker image, and de-identified patient examples.

Results

Patient Overview

Two datasets are studied in this work—the NLST and our VLSP. A total number of 23 505 patients with 64 898 CT scans are included in the NLST (mean age, 62 years; 13 838 men, 9667 women). The VLSP contains a total number of 147 patients with 220 scans (mean age, 65 years; 82 men, 65 women).

Model Performance

The model AUCs and AUPRCs are reported in Table 3. Our method had a higher performance (AUC, 0.878) than the Liao model (20) (image-only) (AUC, 0.864; $P < .05$) and other clinical models including $\text{PLCO}_{\text{M2012}}$ (AUC, 0.692; $P < .05$) and the Brock model (7) (AUC, 0.845; $P < .05$).

External Testing on VLSP

The model AUCs and AUPRCs are shown in Figure 3 and Table 4. Our model was found to have a higher performance (AUC, 0.905) compared with Liao model (20) (AUC, 0.881; $P < .05$) and the Brock model (7) (AUC, 0.782; $P < .05$). These results were found even as our model used different imputed values where the nodule characteristic data elements were missing. In addition, two examples (one cancer, one noncancer) are shown in Figure 4 with predicted cancer probabilities of different methods.

Discussion

Risk prediction modeling can help clinicians make more informed decisions regarding invasive procedures and diagnostic testing. We believe that providing additional information (eg, predicted risk from artificial intelligence) may support subsequent evaluation with physiologic imaging (ie, PET/CT or tissue sampling with biopsy), while a lower risk would support short-term follow-up with CT imaging. All three of these courses are currently acceptable for patients with suspicious pulmonary nodules on screening (ie, Lung-RADS 4). Current guidelines from the American College of Chest Physicians suggest further testing when the risk for lung cancer is 5%–65% and referral for a tissue diagnosis when the risk for cancer is above 65% (34).

The benefit of lung screening is in part owing to annual follow-up, which is recommended for patients with either a Lung-RADS category 1 or category 2 score. These categories are sometimes used interchangeably, particularly for patients with small nodules that remain stable on subsequent examinations. Although small and with low potential of malignancy, these nodules still can be tested and used in developing machine learning models.

Our study demonstrates that a deep learning framework integrating CDEs and CT imaging features can be helpful in lung cancer risk estimation. Risk estimation among lung screening participants will become even more important with the impending expansion of screening guidelines to include those patients who are considered lower risk based only on age and history of tobacco use. However, these patients may be identified as high risk by using machine learning imaging algorithms of low-dose CT scans. For patients with a positive CT scan, CDEs may help the pulmonologist or thoracic surgeon determine the best individualized treatment. Motivated by the synergy between CT and CDEs, we included the CDEs in a multimodal context with CT images for improved risk estimation. Additionally, our model is flexible and can be extended by adding more elements (eg, nodule size), if available. Thus, we believe a potential future direction for improving current standard Lung-RADS recommendations is considering predicted risk from a multimodal deep learning model with standard input of CT images and clinical data.

In this work, we applied established machine learning techniques (eg, multiple instance learning, multimodality fusion)

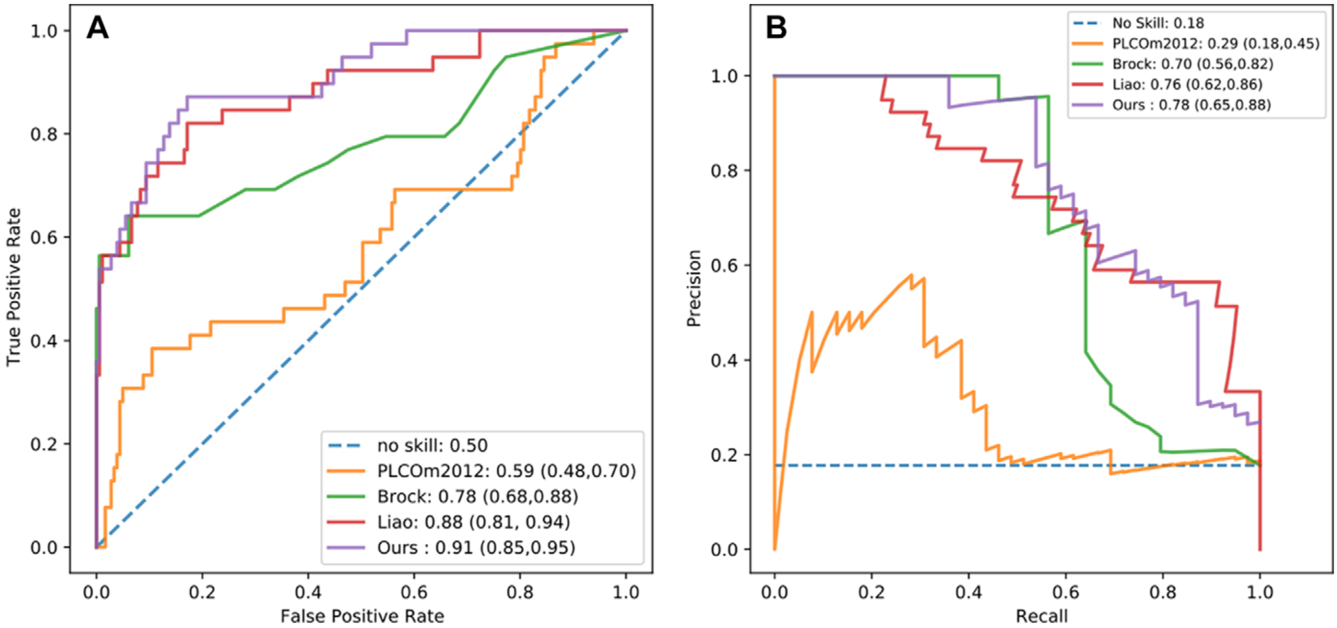

Figure 3: **(A)** Area under the receiver operating characteristic curve (AUC) and **(B)** area under the precision-recall curve (AUPRC) on the external Vanderbilt Lung Screening Program test dataset. **(B)** The AUC and AUPRC of the Brock model were computed by imputing the default values (nodule size, 0 mm for Lung-RADS 1 and 3 mm for Lung-RADS 2; spiculation: no; type: solid; upper lobe achieved by logistic regression) when patient data were not available. Lung-RADS = Lung Reporting and Data System, PLCO = Prostate, Lung, Colorectal, and Ovarian.

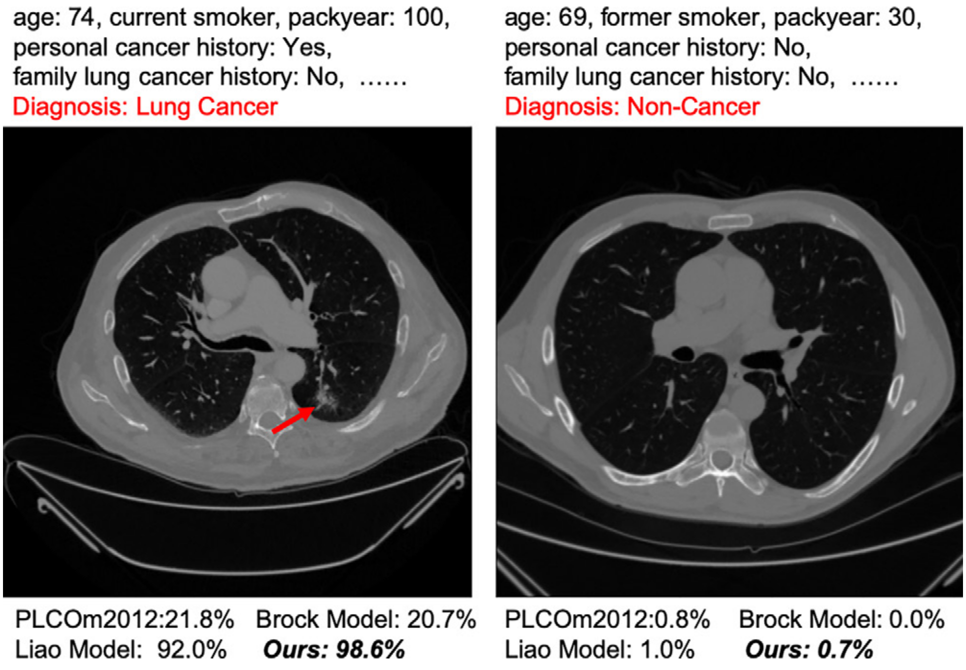
age: 74, current smoker, packyear: 100,
personal cancer history: Yes,
family lung cancer history: No, ……
Diagnosis: Lung Cancer
age: 69, former smoker, packyear: 30,
personal cancer history: No,
family lung cancer history: No, ……
Diagnosis: Non-Cancer
PLCOm2012:21.8% Brock Model: 20.7%
Liao Model: 92.0% Ours: 98.6%
PLCOm2012:0.8% Brock Model: 0.0%
Liao Model: 1.0% Ours: 0.7%

Figure 4: Left: A cancer case. Right: A noncancer case. In each panel, the clinical data are shown above the image and the predicted cancer risks are shown beneath it. Our prediction is calibrated with a sigmoid function. PLCO = Prostate, Lung, Colorectal, and Ovarian.

to improve computer-aided diagnosis. Our approach is automatic to extract high-risk regions (ie, nodule proposals) from CT without radiologist participation and integrate image features with CDEs by deep learning techniques. The false-positive nodules detected are handled by multi-instance learning techniques. Our model is different from radiomics models (eg, Brock model) and Lung-RADS–based evaluation in that our pipeline does not require manual nodule segmentation and nodule characteristics (eg, nodule size) extraction. CDEs are usually collected from clinical decision-making visits and/or questionnaires, which also requires manual effort to collect and input them into systems. The role of the radiologist remains irreplaceable in terms of looking for and reporting clinically significant findings (eg, emphysema, pulmonary fibrosis, atelectasis).

Our model was evaluated by cross-validation with data from the NLST and external testing with an in-house dataset (ie, VLSP). The results show that our model demonstrates higher performance compared with the image-only model and established risk calculators ($P < .05$ for both). The proposed model achieved +2.4% AUC values over the Liao model (image-only model) in external testing using the VLSP dataset; the expected number of patients who would benefit from a better estimate would be 24 out of every 1000 CT scans. Improved discrimination performance for cancer versus noncancer helps inform patient care. We also find that adding more patients for fine-tuning an image-based model (see Appendix E3 [supplement]) can increase prediction performance, which motivates us to use as much data as possible for our multimodal model.

Some other methods, such as the $\mathrm{PLCO}_{\mathrm{M2012}}$, which uses CDEs as inputs for cancer risk estimation and selection criteria for screening, cannot distinguish between cancer and noncancer effectively among the high-risk population. It may be that CDEs only reflect the general "long-term" risk and $\mathrm{PLCO}_{\mathrm{M2012}}$ do not have a data source (eg, CT image) reflecting current symptoms.

Image-only methods (16,20–22) are well developed as a result of the success of feature representation in machine learning (12–15) and large acquisition of medical image data. In lung cancer risk estimation, deep learning methods are especially popular owing to the existence of large, publicly available datasets (eg, Lung Image Database Consortium and Image Database Resource Initiative [LIDC-IDRI] [31], NLST [2]) and the promotion of challenges (eg, Kaggle Data Science Bowl [23], LUNA16 [32]). Li and Fan (16) developed a deep Squeeze-and-Excitation Encoder-Decoder (or DeepSEED), a 3D convolutional neural network with encoder-decoder structure that achieves the state-of-the-art nodule detection performance on LUNA16 and LIDC-IDRI. Liao et al (20) won the Kaggle challenge (23) with a 3D deep neural network that predicts cancer risk by inputting a whole CT image. Ardila et al (21) proposed that a deep learning algorithm using screening CT images can outperform radiologists. The image-based methods are automated, which may be important in a busy imaging practice. These methods ignore that the CDEs can provide complementary information to CT images when predicting cancer risk. Our model integrates the CDEs and CT image features in the end-to-end machine learning framework.

The Brock model incorporates the "current" clinical status of a patient as gleaned from curated lung nodule radiomic features along with the CDEs. Huang et al (5) encoded the human-curated radiographic feature and CDEs into a multilayer perceptron network. The Brock model is widely used in lung cancer risk estimation (eg, McWilliams et al [7]), yet one challenge is that it requires additional human effort to reliably annotate and report nodule features in a structured manner to be used in routine practice. Our model similarly combines the patient-level clinical features (CDEs and radiomic information), but it extracts radiomic features directly from CT images in an efficient and reliably automated manner using a deep learning model.

Our study had several important limitations. We only used a single time point from each patient to predict lung cancer risk; prior CT records may be helpful for prediction. Integrating longitudinal information in this multimodal context will be an interesting topic to explore in the next step. Second, we evaluated the "discrimination" of prediction models with the metrics AUC and AUPRC. We have not yet introduced another aspect, that is, the "uncertainty calibration" (33), in clinical decision making; this concept refers to the agreement between observed probability and predicted risk. In the future we intend to integrate uncertainty calibration in our framework, which would make the machine learning system more reliable. In addition, the currently available data in VLSP is limited compared with popular screening programs, such as the NLST. The potential effect on patient management and ultimately improvement in survival would be better evaluated with larger datasets and including more available CDEs, which we hope to address with future research.

In conclusion, we combined CDEs and CT images for lung cancer risk estimation in a unified machine learning model. Our risk prediction model had a higher performance compared with established risk calculators ($\mathrm{PLCO}_{\mathrm{M2012}}$ [4] and Brock model [7]) and the image-based method (Liao et al [20]) on the NLST and VLSP datasets. Our model can be extended when more data elements are available.

Acknowledgment: We gratefully acknowledge the support of NVIDIA with the donation of the Titan X Pascal GPU used for this research.

Author contributions: Guarantors of integrity of entire study, R.G., Y.T.; 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, R.G., Y.T., M.S.K., K.X., P.P.M., K.L.S., B.A.L.; clinical studies, Y.T., A.B.P., S.D., P.P.M., K.L.S.; experimental studies, R.G., Y.T., M.S.K., Y.H., B.A.L.; statistical analysis, R.G., Y.T., M.S.K., S.D., B.A.L.; and manuscript editing, R.G., Y.T., M.S.K., K.X., A.B.P., Y.H., S.D., P.P.M., K.L.S., B.A.L.

Disclosures of Conflicts of Interest: **R.G.** grant from National Institutes of Health. **Y.T.** No relevant relationships. **M.S.K.** No relevant relationships. **K.X.** No relevant relationships. **A.B.P.** employed by Vanderbilt University Medical Center. **S.S.** No relevant relationships. **Y.H.** grants from NIH and NSF; university employee at Vanderbilt University. **S.D.** grant from National Cancer Institute (U01CA152662). **P.P.M.** No relevant relationships. **K.L.S.** grant from Vanderbilt-Ingram Cancer Center (VICC) (received funding through the Martineau Innovation Award from VICC). **B.A.L.** grant from NIH.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69(1):7–34.
2. National Lung Screening Trial Research Team; Aberle DR, Berg CD, et al. The National Lung Screening Trial: overview and study design. Radiology 2011;258(1):243–253.
3. Humphrey LL, Deffebach M, Pappas M, et al. Screening for lung cancer with low-dose computed tomography: a systematic review to update the US Preventive services task force recommendation. Ann Intern Med 2013;159(6):411–420.
4. Tammemägi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. N Engl J Med 2013;368(8):728–736. [Published correction appears in N Engl J Med 2013;369(4):394.]
5. Huang P, Lin CT, Li Y, et al. Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method. Lancet Digit Health 2019;1(7):e353–e362.
6. Tammemagi MC, Schmidt H, Martel S, et al. Participant selection for lung cancer screening by risk modelling (the Pan-Canadian Early Detection of Lung Cancer [PanCan] study): a single-arm, prospective study. Lancet Oncol 2017;18(11):1523–1531.
7. 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.
8. Henschke CI, Yip R, Yankelevitz DF, Smith JP; International Early Lung Cancer Action Program Investigators*. Definition of a positive test result in computed tomography screening for lung cancer: a cohort study. Ann Intern Med 2013;158(4):246–252.
9. National Lung Screening Trial Research Team; Church TR, Black WC, et al. Results of initial low-dose computed tomographic screening for lung cancer. N Engl J Med 2013;368(21):1980–1991.
10. Horeweg N, van der Aalst CM, Vliegenthart R, et al. Volumetric computed tomography screening for lung cancer: three rounds of the NELSON trial. Eur Respir J 2013;42(6):1659–1667.
11. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.

12. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137–1149.
13. Liu W, Wen Y, Yu Z, Li M, Raj B, Song L. SphereFace: Deep Hypersphere Embedding for Face Recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26, 2017. Piscataway, NJ: IEEE, 2017; 6738–6746.
14. Gao R, Yang F, Yang W, Liao Q. Margin Loss: Making Faces More Separable. IEEE Signal Process Lett 2018;25(2):308–312.
15. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Advances in Neural Information Processing Systems 27, 2014; 2672–2680.
16. Li Y, Fan Y. DeepSEED: 3D Squeeze-and-Excitation Encoder-Decoder Convolutional Neural Networks for Pulmonary Nodule Detection. Proc IEEE Int Symp Biomed Imaging 2020;1866–1869.
17. Shen W, Zhou M, Yang F, Yang C, Tian J. Multi-scale convolutional neural networks for lung nodule classification. Inf Process Med Imaging 2015;24:588–599.
18. Liu L, Dou Q, Chen H, Olatunji IE, Qin J, Heng PA. MTMR-Net: Multi-task Deep Learning with Margin Ranking Loss for Lung Nodule Analysis. In: Stoyanov D, Taylor Z, Carneiro G, et al, eds. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. DLMIA 2018, ML-CDS 2018. Lecture Notes in Computer Science, vol 11045. Cham, Switzerland: Springer, 2018; 74–82.
19. Massion PP, Antic S, Ather S, et al. Assessing the Accuracy of a Deep Learning Method to Risk Stratify Indeterminate Pulmonary Nodules. Am J Respir Crit Care Med 2020;202(2):241–249.
20. 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.
21. 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.]
22. Gao R, Li L, Tang Y, et al. Deep multi-task prediction of lung cancer and cancer-free progression from censored heterogenous clinical imaging. Proc SPIE Int Soc Opt Eng 2020;11313:10.1117/12.2548464.
23. Kaggle. Data Science Bowl 2017. https://www.kaggle.com/c/data-science-bowl-2017. Accessed June 10, 2020.
24. Gao R, Huo Y, Bao S, et al. Distanced LSTM: Time-Distanced Gates in Long Short-Term Memory Models for Lung Cancer Detection. In: Suk HI, Liu M, Yan P, Lian C, eds. Machine Learning in Medical Imaging. MLMI 2019. Lecture Notes in Computer Science, vol 11861. Cham, Switzerland: Springer, 2019; 310–318.
25. Gao R, Huo Y, Bao S, et al. Multi-path x-D recurrent neural networks for collaborative image classification. Neurocomputing 2020;397:48–59.
26. Gao R, Khan MS, Tang Y, et al. Technical Report: Quality Assessment Tool for Machine Learning with Clinical CT. arXiv 2107.12842 [preprint]. https://arxiv.org/abs/2107.12842. Posted July 27, 2021. Accessed July 29, 2021
27. 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 11596, Medical Imaging 2021: Imaging Processing 115961E.
28. Ilse M, Tomczak JM, Welling M. Attention-based Deep Multiple Instance Learning. 35th International Conference on Machine Learning (ICML) 2018.
29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
30. Lung-RADS. Radiology Reference Article. Radiopaedia.org. https://radiopaedia.org/articles/lung-rads?lang=us. Accessed September 26, 2020.
31. Armato SG 3rd, McLennan G, Bidaut L, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 2011;38(2):915–931.
32. Setio AAA, Traverso A, de Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Med Image Anal 2017;42:1–13.
33. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. Proceeding of the 34th International Conference on Machine Learning, PMLR 2017, Vol 70; 1311–1320.
34. Gould MK, Donington J, Lynch WR, et al. Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 2013;143(5 Suppl):e93S–e120S.