Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Lancet Digit Health. Author manuscript; available in PMC 2020 November 01.

Published in final edited form as:

Lancet Digit Health. 2019 November ; 1(7): e353–e362. doi:10.1016/S2589-7500(19)30159-1.

Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method

Peng Huang, PhD[1,17,18], **Cheng T. Lin, MD**[2,18], **Yuliang Li, MS**[3], **Martin C. Tammemagi, PhD**[4], **Malcolm V. Brock, MD**[5], **Sukhinder Atkar-Khattra, BSc**[6], **Yanxun Xu, PhD**[3], **Ping Hu, ScD**[7], **John R. Mayo, MD**[8], **Heidi Schmidt, MD**[9], **Michel Gingras, MD**[10], **Sergio Pasian, MD**[10], **Lori Stewart, MD**[11], **Scott Tsai, MD**[11], **Jean M Seely**[12], **Daria Manos, MD**[13], **Paul Burrowes, MD**[14], **Rick Bhatia, MD**[15], **Ming-Sound Tsao, MD**[9], **Stephen Lam, MD**[16]

[1]Department of Oncology, Johns Hopkins University, Baltimore, Maryland, USA

[2]Department of Radiology, Johns Hopkins University, Baltimore, Maryland, USA

[3]Department of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, Maryland, USA

[4]Department of Community Health Sciences, Brock University, St. Catharines, Ontario, Canada

[5]Department of Surgery, Johns Hopkins University, Baltimore, Maryland, USA

[6]British Columbia Cancer Agency, Vancouver, British Columbia, Canada

[7]Division of Cancer Prevention, National Cancer Institute, Canada

[8]University of British Columbia and Vancouver General Hospital, Vancouver, British Columbia, Canada

[9]University Health Network-Princess Margaret Cancer Centre and Toronto General Hospital, Toronto, Ontario, Canada

[10]Institut universitaire de cardiologie et, de pneumologie de Québec, Canada

Corresponding Author: Peng Huang PhD. Telephone: 1-410-502-0944, Fax: 1-410-955-0859, phuang12@jhmi.edu.

Author contributions

Literature search was done by S.Lam, P.Huang, CT.Lin, MC.Tammemagi, and MV.Brock. Figures and tables were prepared by P.Huang. The study was designed by S.Lam, P.Huang, and CT.Lin. NLST training data were provided by the National Cancer Institute of the United States. PanCan validation data were collected by S.Lam, MC.Tammemagi, S.Atkar-Khattra, JR.Mayo, H.Schmidt, M.Gingras, S.Pasian, L.Stewart, S.Tsai, JM.Seely, D.Manos, P.Burrowes, R.Bhatia, MS.Tsao and other members of the Pan-Canadian Early Detection of Lung Cancer Study group. Data analyses were performed by P.Huang, Y.Li, P.Hu, and Y.Xu. Manuscript was drafted by P.Huang, S.Lam, and CT.Lin. Manuscript was edited by P.Huang, S.Lam, CT.Lin, MC.Tammemagi, Y.Li, MV.Brock, P.Hu, and D.Manos. Manuscript was approved by all authors.

Data sharing

The NLST protocol is available at https://www.acrin.org/6654_protocol.aspx. NLST trial summary, data collected, and data dictionary are available at https://biometry.nci.nih.gov/cdas/nlst/ and https://biometry.nci.nih.gov/cdas/datasets/nlst/. NLST data access request can be made through website https://biometry.nci.nih.gov/cdas/contact/nlst/. The PanCan data summary is available at www.brocku.ca/pancan-lung-screen-study. PanCan data access request will be approved on case by case basis by the PanCan Steering Committee.

Full professors: Stephen Lam, Malcolm Brock, Martin Tammemagi, John R. Mayo, Heidi Schmidt, Ming-Sound Tsao.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[11]Department of Diagnostic Imaging, Juravinski Hospital, Hamilton, Ontario, Canada

[12]Ottawa Hospital Research Institute and the University of Ottawa, Ottawa, Ontario, Canada

[13]Dalhousie University, Halifax, Nova Scotia, Canada

[14]University of Calgary, Foothills Medical Centre, Calgary, Alberta, Canada

[15]Memorial University, Newfoundland, Canada

[16]University of British Columbia-British Columbia Cancer Agency and Vancouver General Hospital, Vancouver, British Columbia, Canada

[17]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA

[18]Co-first authors.

Abstract

Background—Current lung cancer screening guidelines use mean diameter, volume or density of the largest lung nodule in the prior computed tomography (CT) or appearance of new nodule to determine the timing of the next CT. We aimed at developing a more accurate screening protocol by estimating the 3-year lung cancer risk after two screening CTs using deep machine learning (ML) of radiologist CT reading and other universally available clinical information.

Methods—A deep machine learning (ML) algorithm was developed from 25,097 participants who had received at least two CT screenings up to two years apart in the National Lung Screening Trial. Double-blinded validation was performed using 2,294 participants from the Pan-Canadian Early Detection of Lung Cancer Study (PanCan). Performance of ML score to inform lung cancer incidence was compared with Lung-RADS and volume doubling time using time-dependent ROC analysis. Exploratory analysis was performed to identify individuals with aggressive cancers and higher mortality rates.

Findings—In the PanCan validation cohort, ML showed excellent discrimination with a 1-, 2- and 3-year time-dependent AUC values for cancer diagnosis of 0·968±0·013, 0·946±0·013 and 0·899±0·017. Although high ML score cohort included only 10% of the PanCan sample, it identified 94%, 85%, and 71% of incident and interval lung cancers diagnosed within 1, 2, and 3 years, respectively, after the second screening CT. Furthermore, individuals with high ML score had significantly higher mortality rates (HR=16·07, p<0·001) compared to those with lower risk.

Interpretation—ML tool that recognizes patterns in both temporal and spatial changes as well as synergy among changes in nodule and non-nodule features may be used to accurately guide clinical management after the next scheduled repeat screening CT.

Keywords

Lung cancer; screening; Lung-RADS; volume doubling time; deep machine learning; ensemble learning; time-dependent ROC; survival analysis

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

INTRODUCTION

The National Lung Screening Trial (NLST) showed a 20% lung cancer mortality reduction with screening using low-dose computed tomography (LDCT) compared to chest radiography.[1] Recently, the Dutch-Belgium NELSON trial reported a 26% reduction in lung cancer mortality in men and up to a 61% mortality reduction in women with LDCT screening versus no screening.[2] The annual screening recommendation by the US Preventive Services Task Force along with Medicare's coverage of screening LDCT have substantially increased the number of both baseline and follow-up screening LDCT scans.

Management of lung nodules after the next scheduled repeat LDCT is generally determined by the most concerning interval change between the previous two scans, which can be changes in the diameter or volume of solid lung nodule, increase in diameter or density of subsolid nodule, or the size of a new lung nodule. In isolation, these changes may not necessarily reflect the overall risk of developing lung cancer in an individual, especially for those with multiple nodules.[3–7] Current guidelines also do not provide estimates of malignancy risk beyond one year, tumor aggressiveness, or lung cancer specific mortality. An annual repeat screening protocol may result in unnecessary radiation exposure and health care resource utilization with added expense for individuals with low malignancy risk.

Recent advances in statistical modeling techniques, especially in machine learning (ML), are bringing forth a paradigm shift in analyzing complex biomedical data. Machine learning can rapidly process large amount of data and identify complex interactions and associations from high dimensional data to enable earlier and more accurate disease diagnosis.[8–13] There are two major types of ML applications in lung cancer risk prediction. One uses computer-aided diagnosis (CAD) that includes nodule detection, segmentation, and feature extraction.[14–17] The other uses manual defined image region of interest (ROI) with expert crafted image features.[18–20] Recent advances in deep ML technique allows the machine to learn from the input data through multiple layers and automatically derive optimal high-level features without need for human engineered features.[8–10] Since the image processing of entire chest volume is computationally intensive, most studies were restricted to a few pre-specified ROIs without analyzing all identified lung nodules. The work from Google's group is probably the first attempt in large scale to apply deep learning to the entire chest CT dataset.[10] Although the advantages of ML techniques have been well recognized,[9–11] there are major hurdles in translating them to manage screen-detected lung nodules or to determine optimal screening intervals. Issues include sample selection bias, over-fitting due to model selection without blinding the external validation sample, use of non-validated threshold, inadequate statistical analysis that do not adjust for variable censoring times, and a lack of large robust independent validation datasets which can compromise the accuracy and reproducibility of the findings. To address a clinical need for a more accurate lung cancer screening protocol, we developed a deep ML prediction algorithm to estimate the 3-year lung cancer risk and lung cancer specific mortality to guide timing of diagnostic tests and screening interval after the next scheduled repeat screening CT using radiologist CT reading and other universally available clinical information.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

METHODS

Training and Validation Datasets

The datasets for training (NLST) and validation (PanCan) have been reported previously.[1,21] In brief, NLST recruited ever-smokers from 33 screening centers, 55 to 74 years of age, who had a history of at least 30 pack-years smoking history and last smoked within the previous 15 years. The median follow-up was 6·5 years. 25,097 participants received repeat annual LDCT screenings. The PanCan study recruited current and former smokers from 8 screening centers across Canada, 50 to 75 years of age, with a 6-year lung cancer risk of at least 2% as determined by the PanCan model, a precursor to the validated $PLCO_{M2012}$ model.[22] The median follow-up was 5·5 years. Among the 2,350 PanCan participants who received follow-up screenings, 56 individuals were removed because they received investigational screenings before the next scheduled annual screening for suspicious lung nodules. Both datasets included demographics, radiology reports from the most recent follow-up LDCT scan (="S2" scan) within two years from baseline screening and from the most recent prior scan (="S1" scan) from all participants (supplement analysis samples).

Study Oversight

Investigators from three institutions, Johns Hopkins University (JHU), BC Cancer (BCC), and the National Cancer Institute (NCI), jointly conducted this double-blinded study using the PanCan study as a validation cohort.[21] Blinded to the PanCan participant's cancer outcome, the JHU team submitted ML risk scores to the NCI team. Blinded to the JHU team's prediction, BCC provided the verified PanCan participant's cancer outcomes to the NCI team. Both JHU team's risk scores and BCC's cancer outcomes were then locked by the NCI team who compared the JHU team's prediction accuracy to the Lung-RADS[3] and NELSON[5] guidelines. After this evaluation was completed, both JHU team and BCC team were unblinded to the PanCan participants' cancer outcomes and JHU team's ML risk scores.

Deep Machine Learning Algorithm

The endpoint was the duration from follow-up S2 screening date to the last date the person was known to be free of pathologically confirmed lung cancer within three years after S2 scan date. Input predictors are listed in Table 1 and Table S1. Observations were censored at the end of year 3. The 3-year period was selected since our intention was to demonstrate the long-term predictive ability of ML algorithm and determine whether an annual screening interval is appropriate in the lowest-risk group. Limiting the follow-up period to three years also increase the probability that incident cancers within this period could be in-part predicted by CT scan findings. Similar to NLST, only non-calcified nodules with size ≥4mm in PanCan were included. Supplementary Table S1 shows how different radiographic variables used in NLST and PanCan were reconciled. Nodule diameter was computed as the average of the longest diameter and the longest perpendicular diameter as measured on a single transverse image (high-spatial frequency, lung "windows"). Nodule margin was dichotomized to spiculation or no spiculation. Attenuation was replaced by a variable with 4 categories: "ground-glass opacity", "solid", "semi-solid", and "others". Nodule location in PanCan data was converted to the corresponding NLST labels. The American College of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Radiology Lung-RADS category was determined for each screen-detected nodule as described previously.[23] The highest Lung-RADS score for the S2 scan was assigned to each individual.

The diameter of a nodule that resolved at S2 was set to 1mm that was considered to be below the scanner detection limit; Similarly, if a nodule was seen in S2 but not in S1, its diameter in S1 was also set to 1mm. Nodule volume was estimated using $V=\pi d^3/6$ where d is the nodule diameter, and nodule volume doubling time (VDT) in days was calculated for the largest non-calcified nodules from the S2 scan using formula $VDT=\frac{D \log_{10} 2}{3 \log_{10}(d_2/d_1)}$ where d_1 and d_2 are diameters at S1 and S2 screenings respectively, and D is the number of days between S1 and S2 scan dates. For new nodules with d_2>1mm, we set d_1=1mm. Since this VDT formula does not provide finite values for individuals who did not have any or new nodules with d_2>1mm, these individuals were assigned to an artificial VDT value of MaxVDT+1, where MaxVDT is the maximum of all finite VDT values (>36500). The reason is that larger VDT is considered to have lower risk, and such assigned value allowed us to compute AUC and to compare VDT with other predictors using all PanCan individuals. In the NLST training sample, since the same nodule may be labeled differently at different screenings, VDT cannot be calculated for individuals who had at least two nodules from the same lobe. Thus, VDT was only used to evaluate the ML's performance on PanCan validation sample where nodule locations were consistently labeled. Two imaging variables unrelated to nodule characteristics were duration of emphysema and duration of cardiovascular disease; both were calculated by the number of days between the first date of reported abnormality (S1 or S2) and the S2 screening date. A new summary interval change variable, *Nchg*, was created using formula $Nchg=\sum_{(i:\ \text{nodules}\ in\ \text{s2})} C_i \times 365/D$ where the summation is computed over all non-calcified nodules with a size of at least 4mm detected on the S2 screening, and Ci is the number of nodule changes among the following: (1) increase in size, (2) increase in density, (3) new from S1.

The neural network used to develop ML predictor was built using multilayer perceptron (MLP). We used two MLP structures with two hidden layers each. The first one has sizes 5 and 2, and second one has sizes 51 and 8. The cross-entropy loss function with L_2 penalty parameter was used. Weights were optimized using quasi-Newton method and stochastic gradient-based method.[24]. Following Faraggi and Simon's suggestion,[25] rectified linear units function was used in the network prior to the final layer where survival random forest was used.[26] Since data were highly unbalanced, downsampling was used in the ensemble learning. The final output is a normalized continuous score ranging from 0 to 1.

Statistical Methods

This study was designed to analyze the residual survival time since the S2 screening date. Data were collected and analyzed based on conditional distribution that participants had received two CT screenings. The primary analysis was to compare the lung cancer prediction accuracy among three predictors (ML, Lung-RADS, and VDT) in the PanCan validation sample using time-dependent area under the ROC curve (AUC) at cut-off years 1, 2, and 3 after the S2 scan.[27,28] The AUC standard deviation and p-values used to compare two AUCs were computed using inverse probability of censoring weighted estimators[29] in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

500 bootstrap simulations. Secondary analysis compared cancer incidence rates (=one minus Kaplan-Meier estimates) among high-risk and low-risk subgroups. The high-risk subgroups of ML, Lung-RADS and VDT were defined by >0·3, >3 (i.e., 4A/4B/4X), and <400 days respectively. The corresponding low-risk subgroups were defined by <0·1, <3, and >600 days respectively. Missing demographics were imputed using multiple imputation method. Missing nodule data were imputed using the last-observation-carrying-forward method.

An exploratory analysis was performed to compare the time-dependent positive predictive value (PPV) of all predictors under common sensitivity levels using PanCan dataset. The PPV was computed using Bays rule PPV(t)=(time-dependent sensitivity at time t)*(1-S(t)) / P(test positive) where S(t) is the Kaplan-Meier estimate of cancer-free probability at time t=1, 2, or 3 years.

Since mortality data were not used in ML algorithm development, exploratory survival analyses were performed to study whether ML algorithm detected lung cancers that were biologically more aggressive using the NLST dataset only as mortality data was not available in PanCan. Deaths not from lung cancer were censored. ML and Lung-RADS were compared by: (1) the mortality rate within high-risk and low-risk subgroups; (2) the added value of ML to Lung-RADS' high-risk subgroup to detect aggressive cancers; (3) the added value of Lung-RADS to ML's high-risk subgroup to detect aggressive cancers. In all analyses, the hazards ratio (HR) was used to compare subgroups, and two-sided p-values were reported from the log-rank test.

Role of funders

Funders have no role in the study. Huang, Lam, and Aka-Khattra have access to all data. Huang, Lam and Lin were responsible for the final decision to submit the manuscript.

RESULTS

Study Samples

Table 1 summarizes participant's demographics and radiology reports. Compared to the NLST, PanCan participants were older at follow-up S2 screening, were more likely to be current smokers, and had more individuals with ≥4mm nodules at both S1 and S2 screenings. A total of 283 Lung-RADS high-risk nodules were found in 235 PanCan S2 scans.

Machine Learning Predictors

An online web-based tool at https://www.caced.jhu.edu is available to provide ML score and estimated cancer incidence probabilities within 3 years after the S2 scan date.

Primary analysis

In the NLST training sample, time-dependent AUC ± standard deviation values of ML are 0·99±0·003, 0·985±0·005, and 0·983±0·005 at years one, two, and three, respectively (Figure 1). They are higher than the corresponding AUC values 0·909±0·019, 0·856±0·022, and 0·811±0·024 from Lung-RADS (all p<0·001). The blinded ML prediction was originally

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

performed on all 2,350 PanCan individuals. But final analysis excluded 56 individuals who received investigational screenings (see supplement appendix for details). Final AUC values from ML were 0·968±0·013, 0·946±0·013, and 0·899±0·017 (Figure 1). They are higher than the updated Lung-RADS AUC values of 0·944±0·016, 0·908±0·019, and 0·858±0·022 (p=0·202, 0·048, 0·028); and VDT AUC values of 0·830±0·030, 0·777±0·029, and 0·762±0·026 (all p<0·001).

Secondary analysis

Using Lung-RADS, 5%(1,333/25,097) of the NLST participants were classified as high-risk (4A/4B/4X) compares to 2%(392/25,097) using ML>0·3 criterion (Table 2 and Figure 2). Among those that were deemed high-risk, Lung-RADS identified 77%(274/358) and 66%(308/464) of all lung cancers diagnosed within 12 and 24 months respectively. The corresponding figures using ML were 87% (313/358) and 74·1%(344/464) (Table 2). In the PanCan study, 10%(235/2,294), 10%(221/2,294) and 8%(192/2,294) would be deemed high-risk by Lung-RADS, ML and VDT respectively. Among those that were deemed high-risk, Lung-RADS identified 89%(48/54) and 81%(58/72) of all lung cancers diagnosed within 12 and 24 months respectively. The corresponding figures were 94%(51/54) and 85%(61/72) with ML, and 57%(31/54) and 50%(36/72) with VDT (Table 2).

Using Lung-RADS, 93% (23,458/25,097) of the NLST participants were classified as low-risk (1 or 2). With ML≤0·1, 97%(24,234/25,097) are classified as low-risk. The proportion of low risk participants with lung cancer diagnosed within 24 months are 0·62%(145/23,458) and 0·21%(51/24,234) for Lung-RADS and ML≤0·1 respectively. For PanCan, the proportion of participants classified as low-risk would be 74%(1,698/2,294), 55%(1,266/2,294) and 90%(2,075/2,294) for Lung-RADS, ML≤0.1 and VDT respectively. The proportion of low-risk individuals diagnosed with lung cancer within 24 months are 0·53%(9/1,698), 0·16%(2/1,266) and 1·49%(31/2,075) respectively (Table 2).

Exploratory analyses

When the common sensitivity level was set between 0·85 and 0·90, using all 1,070 PanCan individuals who had at least one non-calcified nodule with size 4mm or above in their S2 scans, ML has higher time-dependent PPV than Lung-RADS and VDT in all three years (Figure S1).

Among all 353 and 327 lung cancers included in ML and Lung-RADS' high-risk subgroups, 231 and 206 of them were in pathological stage I. Higher lung cancer mortality rates were observed from ML high-risk subgroups as compared to the Lung-RADS high-risk subgroup (supplementary Figure S2). ML low-risk subgroup had uniformly lower mortality rate compared to Lung-RADS low-risk subgroups. Within both Lung-RADS=4A/4B/4X and 4B subgroups, individuals with ML high-risk scores also had significantly higher mortality rates (HR=16·07 and 31·79, both p<0·001, Figure 3, A and B). However, Lung-RADS could not stratify the mortality rate within the ML high-risk subgroup (Figure 3, C and D).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

DISCUSSION

We developed a deep learning algorithm that accounted for all relevant nodule and non-nodule features on the screening chest CTs and accurately predicted the presence of lung cancer within a three-year period. The ML algorithm maintained high accuracy and was generalizable to an external dataset from another country. A double blinded experimental design was used to avoid multiple fitting attempts which could lead to non-reproducible and deceptively high accuracy. Compared to Lung-RADS, ML classified a high-risk group that was smaller and had a higher proportion of cancers (Table 2). ML also identified more accurately those with very low risk of lung cancer within 2 years (0·16%, compares to 1·49% with VDT and 0·53% with Lung-RADS). The appropriateness of a biennial repeat screening protocol for very low risk participants requires further prospective validation studies.

To understand what information has been learnt by the ML algorithm, we have also extracted data from the first hidden layer of the ML algorithm to construct an explicit formula ML1 (Table S2). We found the most important features are related to the changes from S1 to S2, such as *Nchg*. Unlike Lung-RADS, the ML algorithm combined both temporal and spatial changes from S1 to S2 screenings in all nodules to improve cancer prediction. As illustrated in Figure S3, the presence of incidence nodules is more important than the nodule size measured at any single time point.

While mortality data was not a training variable in ML development, ML demonstrated added value in stratifying lung cancer mortality risk (Figure 3). In the NLST cohort, ML outperformed Lung-RADS in predicting deaths from lung cancer. Among participants deemed high-risk by Lung-RADS criteria, ML could further distinguish the subgroup with significantly higher risk of lung cancer mortality (Figure 3). Similar mortality data is not currently available from PanCan patients.

Our study was designed to avoid some common pitfalls in previous ML studies. Although a case-control design is helpful to demonstrate ML's added value[11,19,20], the study sample from such a design would not be representative of the general screening population and therefore its ML predictor could not be directly applied to clinical practice. Most ML studies did not blind the validation sample's cancer outcomes. This could result in multiple attempts of fitting spurious associations that could lead to an artificially crafted highly accurate model. Many published ML studies did not differentiate cancers diagnosed at different time points in both ML algorithm development and its AUC evaluation.[17,19,20] Our study outcomes were derived from survival analysis that takes an individual's length of follow-up time into consideration. This approach not only reduces the bias from different follow-up lengths of participants, but also associates higher ML risk scores with earlier lung cancer diagnosis time. Early censored individuals were not treated as non-cancers in later years because these individuals could also develop cancers later. This is a critical difference between our approach and most existing methods.[10,19] If this censoring was not adjusted, the sensitivity/specificity of the three predictors in PanCan validation sample would be 80%/76% (Lung-RADS), 88%/60% (ML), and 54%/91% (VDT) respectively. The logistic

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

regression based AUC=90% from ML was significantly higher than AUC=86% (p=0·032) from Lung-RADS and AUC=76.% from VDT (p<0·001).

In contrast to currently available malignancy risk prediction tools or guidelines that are nodule-based which quantify an individual's malignancy risk using the largest dominant nodule,[3–6,21,23,30] ML takes into account the aggregate changes in nodule characteristics and non-nodule features for each individual. Examining the potential interactions between different nodules is important since 39% (2,852/7,307) of NLST individuals and 50% (532/1,070) of PanCan individuals had at least two non-calcified nodules with size ≥4mm on their S2 scans. Basing guidelines on the largest nodule can be problematic since the PanCan study has previously shown that 20% of malignancy arose from non-dominant nodules.[7] Although Google's recent work has incorporated image features from multiple nodules,[10] its prediction was limited to one-year cancer risk and not evaluated from the survival analysis. An additional strength of our study is the ability to recognize patterns in both temporal and spatial changes including the synergy among changes from different nodules. Our algorithm can be used to provide guidance to clinicians, to personalize the repeat screening interval (Figures 2 and Table 2), and to determine the urgency for diagnostic investigations to rule out lung cancer (Figure 3), in a manner that is not currently available in existing clinical practice guidelines.

Several limitations should be considered in this study. First, the LDCT imaging features were interpreted by experienced chest radiologists in large academic centers. Less stringent quality assurance in CT reporting may not produce the same results. Whether the use of CAD or radiomic features can improve the accuracy of our ML algorithm is the subject of future studies. Secondly, nodule size was calculated by averaging its long- and short-axis diameter on one axial image slice. The VDT was calculated from the mean diameter instead of volumetric measurement. The simplicity of using universally available radiologist's reading and other clinical information makes our method applicable to a broader setting where image processing tools are not available. Thirdly, setting the threshold for ML high-risk subgroup at 0·3 was arbitrarily (albeit blindly) chosen to balance the sample size in the ML high-risk subgroup and its prediction accuracy based on the NLST training sample. It may not be the optimal threshold value. Fourthly, individuals in our cohort on average underwent annual screening LDCTs. While our methodology allows for extrapolation of risk assessment in participants who had shorter (e.g. 3 or 6 months) or longer than 12 months interval between the last two screening CTs, additional validation studies are needed. Lastly, as a consequence of developing a "black-box" ML algorithm, the algorithm may not be easily reproduced and analyzed by others. The website: https://www.caced.jhu.edu was created to facilitate other investigators to obtain and further evaluate our ML algorithm using their own data. The interface of this website was designed to illustrate the variables that need to be entered in order to derive the risk score. In practice, the information can be automatically populated by the computer to generate the malignancy risk score. In institutions where structured reporting templates are utilized, it is possible to automate data extraction using natural language processing so long as the required data elements are recorded in radiology reports.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To our knowledge, this is the first accurate and externally validated ML tool with practical application to guide clinicians in lung cancer screening programs. Our study provides the framework to prospectively evaluate different screening intervals and more urgent diagnostic approach for suspicious lung nodules based on malignancy risk.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

We want to thank Dr. Paul Pinsky of the National Cancer Institute (NCI) for providing blinding of the ML scores and the clinical outcome of the validation study, NCI for access to the NCI data collected by the NLST, and Johns Hopkins University Whiting School of Engineering IT group in providing support and maintenance of our web-based machine learning tool.

Funding

This study was funded by Allegheny Health Network, Johns Hopkins University Discovery Award, P30CA006973, the Terry Fox Research Institute, and the BC Cancer Foundation.

Declaration of interests

P.Huang received fund from Allegheny Health Network. S.Lam, S.Atkar-Khattra, MC.Tammemagi, S.Atkar-Khattra, JR.Mayo, H.Schmidt, M.Gingras, S.Pasian, L.Stewart, S.Tsai, JM.Seely, D.Manos, P.Burrowes, R.Bhatia, and MS.Tsao received funds from the Terry Fox Research Institute and the BC Cancer Foundation.

References

1. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365(5):395–409. [PubMed: 21714641]
2. De Koning H, Van Der Aalst C, Ten Haaf K, Oudkerk M Effects of Volume CT Lung Cancer Screening: Mortality Results of the NELSON Randomised-Controlled Population Based Trial. Journal of Thoracic Oncology. 2018;13(10):S185.
3. American College of Radiology. Lung CT Screening Reporting & Data System June 23, 2017 [Available from: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads.
4. Wood DE, Kazerooni EA, Baum SL, Eapen GA, Ettinger DS, Hou L, et al. Lung Cancer Screening, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw. 2018;16(4):412–41. [PubMed: 29632061]
5. Oudkerk M, Devaraj A, Vliegenthart R, Henzler T, Prosch H, Heussel CP, et al. European position statement on lung cancer screening. Lancet Oncol. 2017;18(12):e754–e66. [PubMed: 29208441]
6. Pinsky PF, Gierada DS, Nath PH, Munden R. Lung Cancer Risk Associated With New Solid Nodules in the National Lung Screening Trial. AJR Am J Roentgenol. 2017;209(5):1009–14. [PubMed: 28898131]
7. McWilliams A, Tammemagi MC, Mayo JR, Roberts H, Liu G, Soghrati K, et al. Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med. 2013;369(10):910–9. [PubMed: 24004118]
8. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. [PubMed: 28778026]
9. van Ginneken B Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. Radiol Phys Technol. 2017;10(1):23–32. [PubMed: 28211015]
10. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

11. Huang P, Park S, Yan R, Lee J, Chu LC, Lin CT, et al. Added Value of Computer-aided CT Image Features for Early Lung Cancer Diagnosis with Small Pulmonary Nodules: A Matched Case-Control Study. Radiology. 2017:162725.
12. Ciompi F, Chung K, van Riel SJ, Setio AAA, Gerke PK, Jacobs C, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Sci Rep. 2017;7:46479. [PubMed: 28422152]
13. Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Ther. 2015;8:2015–22. [PubMed: 26346558]
14. El-Baz A, Beache GM, Gimel'farb G, Suzuki K, Okada K, Elnakib A, et al. Computer-aided diagnosis systems for lung cancer: challenges and methodologies. International journal of biomedical imaging. 2013;2013.
15. Huang P, Park S, Yan R, Lee J, Chu L, Cheng T, et al. Lung cancer early diagnosis through image texture analysis and machine learning. Machine Intelligence in Medical Imaging 2nd Society for Imaging Informatics in Medicine Conference 2017;Baltimore, Maryland(September 26–27, 2017.).
16. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006. [PubMed: 24892406]
17. Peikert T, Duan F, Rajagopalan S, Karwoski RA, Clay R, Robb RA, et al. Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the National Lung Screening Trial. PLoS One. 2018;13(5):e0196910. [PubMed: 29758038]
18. Ganeshan B, Abaleke S, Young RC, Chatwin CR, Miles KA. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. Cancer Imaging. 2010;10:137–43. [PubMed: 20605762]
19. Cherezov D, Hawkins SH, Goldgof DB, Hall LO, Liu Y, Li Q, et al. Delta radiomic features improve prediction for lung cancer incidence: A nested case-control analysis of the National Lung Screening Trial. Cancer Med. 2018;7(12):6340–56. [PubMed: 30507033]
20. Paul R, Hawkins SH, Schabath MB, Gillies RJ, Hall LO, Goldgof DB. Predicting malignant nodules by fusing deep features with classical radiomics features. J Med Imaging (Bellingham). 2018;5(1):011021. [PubMed: 29594181]
21. Tammemagi MC, Schmidt H, Martel S, McWilliams A, Goffin JR, Johnston MR, 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):1523–31. [PubMed: 29055736]
22. Tammemagi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection criteria for lung-cancer screening. N Engl J Med. 2013;368(8):728–36. [PubMed: 23425165]
23. Pinsky PF, Gierada DS, Black W, Munden R, Nath H, Aberle D, et al. Performance of Lung-RADS in the National Lung Screening Trial: a retrospective assessment. Ann Intern Med. 2015;162(7):485–91. [PubMed: 25664444]
24. Kingma DP, Ba JL. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980: arXiv [cs.LG]:1412.6980; 2014.
25. Faraggi D, Simon R. A neural network model for survival data. Stat Med. 1995;14(1):73–82. [PubMed: 7701159]
26. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random Survival Forests. Ann Appl Stat. 2008;2(3):841–60.
27. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000;56(2):337–44. [PubMed: 10877287]
28. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005;61(1):92–105. [PubMed: 15737082]
29. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med. 2013;32(30):5381–97. [PubMed: 24027076]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

30. Callister ME, Baldwin DR, Akram AR, Barnard S, Cane P, Draffan J, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. Thorax. 2015;70 Suppl 2:ii1–ii54. [PubMed: 26082159]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Research in context

Evidence before this study

We searched PubMed, Medline, and the Cochrane Library from January 1, 1980 to March 31, 2019 using combinations of words or terms that included: “lung” or “pulmonary”, “cancer” or “neoplasm”, “screening” or “early detection”, “Radiomics”, “deep machine learning”, and “computer-aided diagnosis”. Previous lung cancer screening studies used a few (mostly one or two) prespecified lung regions of interest to predict a person’s lung cancer risk score except the paper by Ardila et al (Nature Medicine, 2019). None of the studies has used a double-blinded design in independent validation sample prediction, adjusted potential bias due to individual’s variable length of follow up (or censoring) time in both prediction algorithm development and prediction accuracy evaluation, provided estimates of malignancy risk beyond one year, tumor aggressiveness, or lung cancer specific mortality. Management of lung nodules after the next scheduled repeat screening computed tomography is based on changes in the diameter or volume of solid lung nodule, increase in diameter or density of subsolid nodule, or the size of a new lung nodule, which may not necessarily reflect the overall risk of developing lung cancer in an individual especially those with multiple nodules. Nor do these guidelines provide estimates of malignancy risk beyond one year, tumor aggressiveness, or lung cancer specific mortality.

Added value of this study

Current lung cancer screening guidelines and nodule management protocols recommend regular annual repeat CT, early recall imaging studies, or triage to a diagnostic pathway depending on the estimated malignancy risk. Our study is the first to develop a deep machine learning prediction algorithm using universally available nodule and non-nodule features without computer-aided diagnostic tools to estimate a person’s 3-year lung cancer risk and associated lung cancer-specific mortality. The deep learning algorithm identifies 10% of the screening population who may benefit from prompt diagnostic workup for biologically aggressive tumor on the one hand, and the 55% of individuals with a very low 2-year malignancy risk of 0·16% who can safely undergo the next scheduled screening CT in two years instead of annually. Our study addresses shortcoming in previous deep learning studies by using a large unselected dataset to avoid bias and over-fitting, adequate statistical models that combine information from multiple nodule and non-nodule associated lung abnormalities with adjustment for variable censoring times, and blinded validation in a well-annotated external dataset. Our algorithm can be used to provide guidance to clinicians, to personalize the repeat screening interval, and to determine the urgency for diagnostic investigations to rule out lung cancer, in a manner that is not currently available in existing clinical practice guidelines.

Implications of all the available evidence

Our study demonstrates that readily available clinical and radiologist-interpreted CT information can be mined using deep machine learning to personalize the repeat screening interval and to determine the urgency for diagnostic investigations to rule out

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

lung cancer. The added value of automated computer image analysis of CT scans needs to be compared with what can be achieved with radiologist reading alone.

At a Glance Commentary

The current study demonstrated that a deep machine learning algorithm that recognizes patterns in both temporal and spatial changes as well as synergy among changes in nodule and non-nodule features can identify individuals with higher lung cancer risk and higher lung cancer mortality. A risk based personalized strategy using an accurate lung cancer risk prediction model is expected to improve the effectiveness of lung cancer screening programs.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

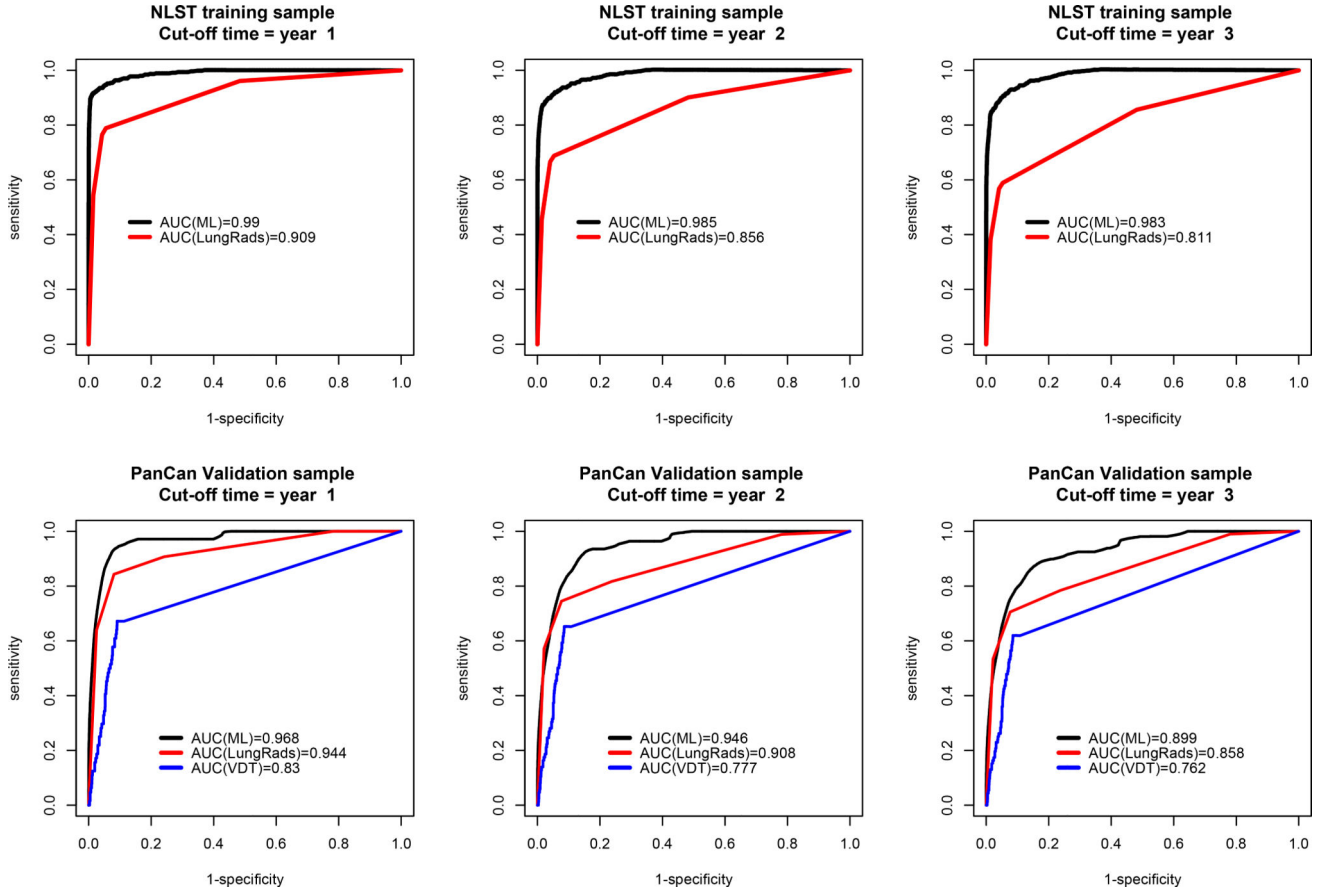

Figure 1.

Comparison of the area under the time dependent receiver operating characteristic curve (AUC) at cut-off time year = 1, 2, and 3 among two risk predictors: ML and Lung-RADS in the NLST training (N=25,097) samples, and three risk predictors, including volume doubling time (VDT), in the PanCan validation (N=2,294) samples. The VDT cannot be reliably estimated in the NLST sample. In the NLST, the p-values of AUC difference between ML and Lung-RADS were <0.001 in all three years. In PanCan, the p-values of AUC difference between ML and Lung-RADS were 0·202, 0·048, and 0·028 for years 1, 2, and 3 respectively; the p-values of AUC difference between ML and VDT were <0.001 for all three years.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

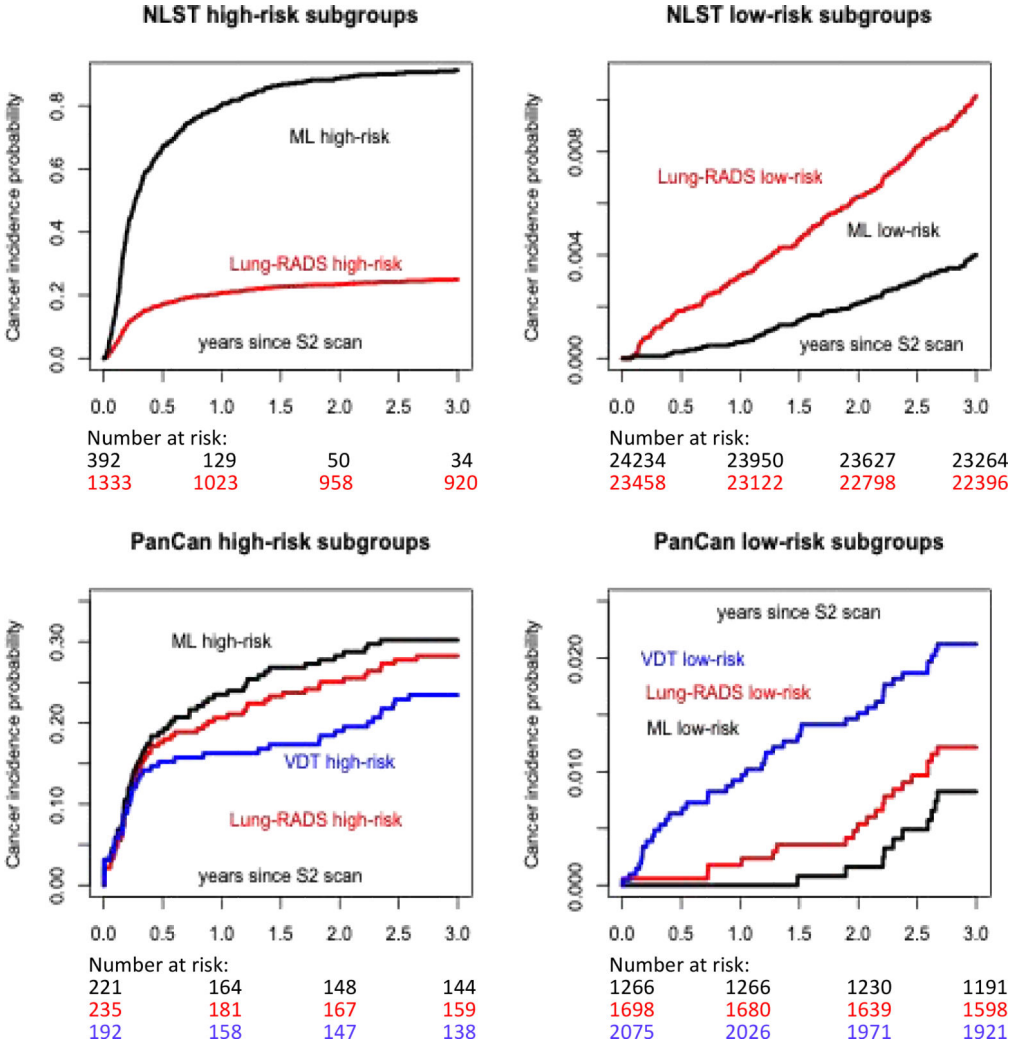

Figure 2.

Comparison of high-risk and low-risk subgroups defined by ML and Lung-RADS for the NLST training sample, and three risk predictors, including volume doubling time (VDT), for the PanCan validation sample. The VDT cannot be reliably estimated in the NLST sample. The cancer incidence probability was computed as one minus Kaplan-Meier estimate of cancer-free survival probability estimate.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

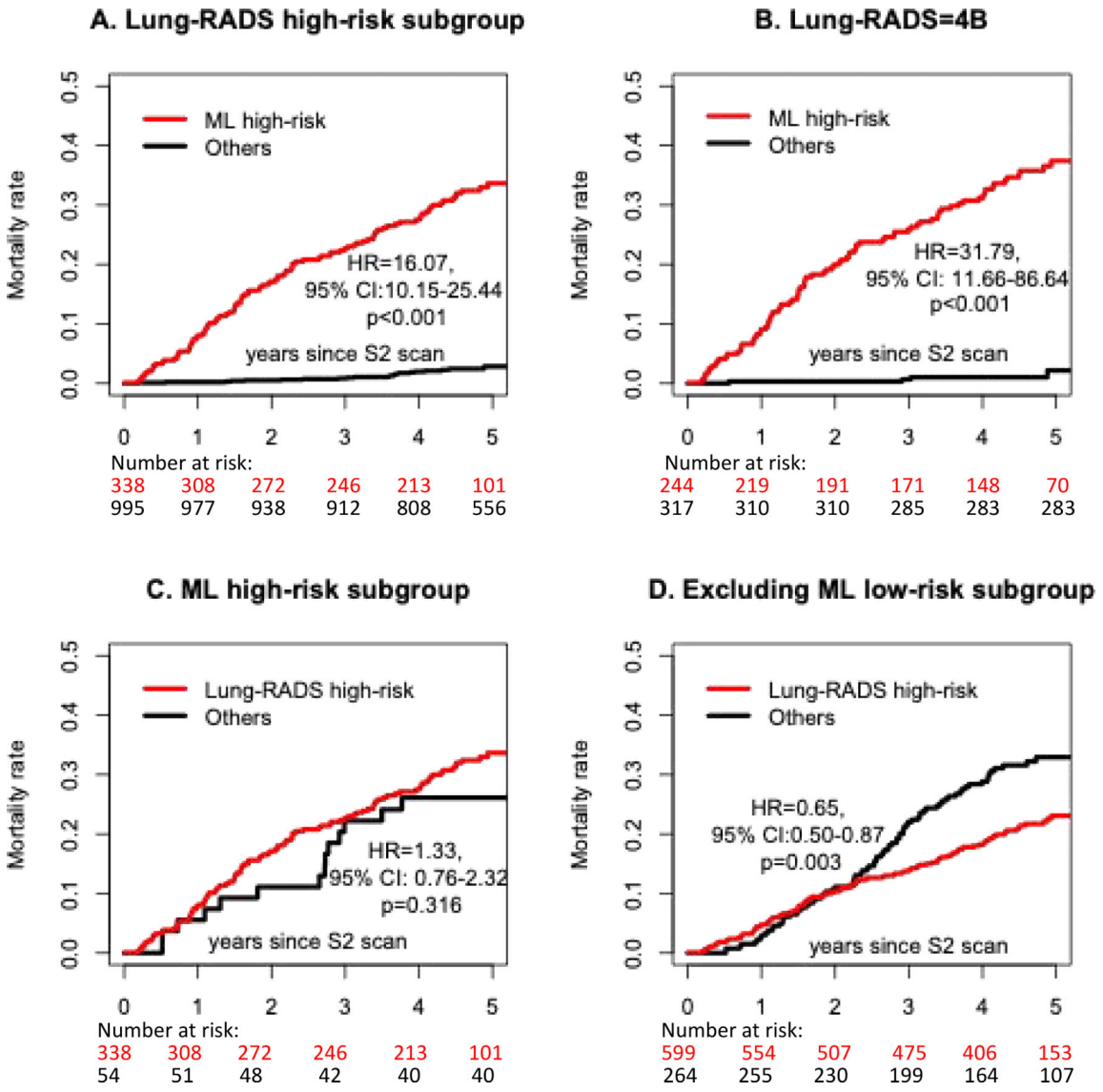

Figure 3.

Individuals with high ML scores had higher lung cancer associated mortality rate within the Lung-RADS high-risk subgroups (A and B). However, Lung-RADS did not separate lung cancer mortality risks within ML high-risk subgroup (C). When excluding ML low-risk individuals, the Lung-RADS high-risk subgroup had an even lower lung cancer mortality rate than the Lung-RADS low-risk subgroup (D). HR = hazards ratio, CI = confidence interval.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1.

Demographic, radiological and clinical outcome data

	NLST	Pan Can	p value
Total number of participants	25,097	2,294	
Age at S2 screening (mean±standard deviation)	62·4±5·0	64·3±5·9	< 0·001
Number of females (%)	10,281 (41%)	1,028 (45%)	0·0004
Current smoker (%)	11,905 (47%)	1,406 (61%)	< 0·001
Pack-years	55·8±23·8	53·9±23·3	0·0002
Duration of smoking (years)	39·7±7·3	44·1±5·9	< 0·001
Age started smoking (years)	16·7±3·7	16·0±3·2	< 0·001
Average number of cigarettes per day	28·4±11·4	24·7±10·7	< 0·001
Family history of lung cancer (%)	5,460 (22%)	757 (33%)	< 0·001
Emphysema (%)	9,048 (36%)	759 (33%)	0·005
Moderate or severe calcification left anterior descending artery, right circumflex, or right coronary artery at baseline		663 (29%)	
Significant cardiovascular abnormality	1,500 (6%)		
S1 screening			
Number of participants with non-calcified nodule at least 4 mm	6,989 (28%)	1,099 (48%)	< 0·001
Number of participants with 2 or more non-calcified nodule at least 4 mm	2,631	520	
Total number of non-calcified nodules at least 4 mm	11,763	2,236	
Largest non-calcified nodule size (mean±standard deviation) among participants with non-calcified nodules. One largest nodule per person	6·7±4·5	6·6±3·4	0·0834
S2 screening			
Number of participants with non-calcified nodule at least 4 mm	7,307 (29%)	1,070 (47%)	< 0·001
Number of participants with 2 or more non-calcified nodule at least 4mm	2,852	532	
Total number of non-calcified nodules at least 4mm	12,602	2,229	
Total number of participants with Lung-RADS high-risk (4A/4B/4X) nodules	1,333	235	
Largest non-calcified nodule size (mean±standard deviation) among participants with non-calcified nodules. One largest nodule per person	6·89±5·57	6·79±3·90	0·2994
Number of days (mean±standard deviation) between S1 and S2 screenings	374·85±79·56	360·04±128·03	
Number of cancers diagnosed in year 1 after S2 screening date	358	54	
Number of cancers diagnosed in year 2 after S2 screening date	106	18	
Number of cancers diagnosed in year 3 after of S2 screening date	110	20	
Number of participants censored in year 1 after S2 screening date	303	37	
Number of participants censored in year 2 after S2 screening date	294	47	
Number of participants censored in year 3 after S2 screening date	340	60	

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2.

Cumulated number of cancers in different high and low risk subgroups at 6-, 12- and 24-months follow up visits after the S2 screening date

		Number of individuals	Cumulative number of cancers diagnosed at follow up visit			
			6 months	12 months	24 months	36 months
NLST training sample						
Total screened		25097	274	358	464	574
High-risk subgroup	Lung-RADS=4B	561	170	194	212	221
	Lung-RADS= 4A/4B/4X	1333	227	274	308	327
	ML>0.3	392	260	313	344	353
Low-risk subgroup	Lung-RADS=1 or 2	23458	43	75	145	234
	ML ≤ 0.1	24234	6	15	51	95
PanCan validation sample						
Total screened		2294	43	54	72	92
High-risk subgroup	Lung-RADS=4B	97	35	38	44	49
	Lung-RADS=4A/4B/4X	235	41	48	58	65
	VDT < 400 days	192	29	31	36	44
	ML>0.3	221	41	51	61	65
Low-risk subgroup	Lung-RADS=1 or 2	1698	1	3	9	20
	ML ≤ 0.1	1266	0	0	2	10
	VDT >600 days	2075	13	19	31	43