RESEARCH ARTICLE

A multi-parameterized artificial neural network for lung cancer risk prediction

Gregory R. Hart☯, David A. Roffman☯¤, Roy Decker, Jun Deng*

Department of Therapeutic Radiology, School of Medicine, Yale University, New Haven, Connecticut, United States of America

☯ These authors contributed equally to this work.
¤ Current address: Sun Nuclear Corporation, Melbourne, Florida, United States of America
* jun.deng@yale.edu

Citation: Hart GR, Roffman DA, Decker R, Deng J (2018) A multi-parameterized artificial neural network for lung cancer risk prediction. PLoS ONE 13(10): e0205264. https://doi.org/10.1371/journal.pone.0205264

Editor: Yuchen Qiu, UNITED STATES

Received: January 26, 2018

Accepted: September 21, 2018

Published: October 24, 2018

Data Availability Statement: All the data used in this study are publicly available from the CDC through the following URL: https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm. The authors did not have any special access privileges and those interested would be able to access the data in the same manner as the authors.

Funding: Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB022589. The content is solely the responsibility of the authors and does not

Abstract

The objective of this study is to train and validate a multi-parameterized artificial neural network (ANN) based on personal health information to predict lung cancer risk with high sensitivity and specificity. The 1997-2015 National Health Interview Survey adult data was used to train and validate our ANN, with inputs: gender, age, BMI, diabetes, smoking status, emphysema, asthma, race, Hispanic ethnicity, hypertension, heart diseases, vigorous exercise habits, and history of stroke. We identified 648 cancer and 488,418 non-cancer cases. For the training set the sensitivity was 79.8% (95% CI, 75.9%-83.6%), specificity was 79.9% (79.8%-80.1%), and AUC was 0.86 (0.85-0.88). For the validation set sensitivity was 75.3% (68.9%-81.6%), specificity was 80.6% (80.3%-80.8%), and AUC was 0.86 (0.84-0.89). Our results indicate that the use of an ANN based on personal health information gives high specificity and modest sensitivity for lung cancer detection, offering a cost-effective and non-invasive clinical tool for risk stratification.

Introduction

Approximately 14% of new cancer cases each year in the United States are lung cancer, but the number of deaths related to lung cancer exceed those from breast, prostate, and colon cancers combined [1]. Even though it is well documented that smoking is the main causal factor, a predictive model that incorporates the synergetic effects of a multitude of patient-related factors and other health information would be useful in the evaluation of persons perceived to be a risk. In this work, we assess the aptitude of such a model developed from training an artificial neural network with the National Health Interview Survey (NHIS) datasets.

There are three main types of lung cancer: non-small cell lung cancer (about 85%), small cell lung cancer (about 10-15%), and lung carcinoid tumors (fewer than 5%) [2]. The standard method of detection by screening is low-dose computed tomography (LDCT) [3]. However, the Centers for Disease Control and Prevention (CDC) indicate that repeated exposure to low dose radiation increases cancer risk. The United States Preventive Services Task Force (USPSTF) recommends screening only for those who have 30 pack years or more of smoking and are current smokers or have stopped within the last 15 years, and are 55-80 years old [4]. While smoking is the primary risk factor, there are other relevant factors such as a family

necessarily represent the official views of the National Institutes of Health. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

history of cancer, diet, and exposure to environmental tobacco smoke (second-hand smoke), radon, asbestos, or other carcinogens [5]. Lung cancers detected at a local stage has a 55% 5-year survival rate; however the majority of lung cancer patients are diagnosed with more advanced disease, with much lower survival rates (overall 5-year survival rate of 18%) [6].

The high risk population identified by the USPSTF lung cancer screening criteria is estimated to include approximately 8—9 million individuals in the United States. While the USPSTF recommendation represents an enormous step forward in early detection for lung cancer, there is ongoing debate as to whether the criteria include individuals whose risk is not high enough to warrant screening [7] and conversely exclude other individuals whose risk is demonstrably high by modeling studies [8, 9]. There has been intense interest in developing methods to more accurately identify individuals at high risk for lung cancer that incorporate demographic as well as biologic inputs. Accordingly many models have been created using a variety of methods such as logistic regression [10–12], restricted cubic splines [8, 13], and two-stage clonal expansion models [14, 15]. These methods have had varying success with AUCs of 0.57-0.88 with the average a little over 0.7 [16]. To the best of our knowledge, our work is the first study that uses machine learning algorithms on this type of data to predict lung cancer risk.

The aim of this study is to investigate a novel approach in predicting lung cancer risk, using a multi-parameterized artificial neural network (ANN) based on personal health information extracted from the National Health Interview Survey (NHIS) datasets. We hypothesized that a multi-parameterized ANN model using readily available clinical and demographic information commonly found in the electronic medical record (EMR) systems would be an effective clinical tool to predict and stratify lung cancer risk for individuals.

Materials and methods

Datasets and patient selection

We obtained NHIS adult survey files, from the CDC website, related to clinical and demographic status, including the corresponding manuals and criteria, which vary by year [17]. We used the NHIS survey datasets from 1997-2015, with the exception of 2004 due to known inaccuracies in the data file. The response rate for the NHIS adult survey is about 80% and we can only view the data that has been collected and filtered by the NHIS [18].

The USPSTF criteria for lung cancer screening guidelines are well defined. However, there is ongoing discussion regarding groups at high risk that are identifiable by modeling but are currently excluded from screening as they do not fit the USPSTF guideline [8, 9]. We decided to include the entire NHIS adult population on the basis of including as many cases of lung cancer as possible. We selected model inputs based on known or putative lung cancer risks factors, as well as clinical and demographic information in the dataset: age [1], body mass index (BMI) [19], diabetic status [20], smoking status, emphysema [21], asthma [22], race [23], Hispanic ethnicity [23], hypertension [24], heart diseases, vigorous exercise habits [25], and history of stroke [26]. The demographics of the entire sample used are shown in Table 1.

We used 70% of the data (454 lung cancer cases and 341,893 never cancer cases) for training and 30% for validation (195 lung cancer cases and 146,524 never cancer cases) with the selection being randomized for each group. Lung cancer cases meeting the inclusion criteria were limited to patients with lung cancer as the first diagnosed malignancy that occurred within 4 years of the survey date. Several of the inputs for our ANN are time-dependent, such as BMI and diabetic status. We selected a four-year cutoff as a compromise between the time-dependent aspects of the problem and the sample size restriction required for training and validation. Note that this four-year cutoff only applies to the lung cancer cases. The prevalence of cancer in our training and testing set is about twice the annual incidence rate [6]

Table 1. The demographics of the NHIS dataset that was used in our ANN. We show means and standard deviations for the continuous variables, means for the binary variables, and the percentage for each race.

Input	Lung Cancer	Non-Cancer
Age	65.6 (±11.8)	46.1 (±17.6)
BMI	25.8 (±5.9)	27.3 (±6.0)
Heart Disease Score	0.13 (±0.22)	0.040 (±0.13)
Number of Vigorous Exercise done per week	0.38 (±1.7)	1.60 (3.0)
Female	53.8%	54.9%
Ever Smoked	83.8%	41.8%
Has Emphysema	24.1%	1.53%
Has Asthma	18.8%	11.2%
Has Diabetes	17.4%	7.92%
Ever Had a Stroke	9.57%	2.55%
Has Hypertension	18.8%	11.2%
Hispanic Ethnicity	7.10%	16.7%
Race:		
Caucasian	82.4%	77.3%
African American	14.2%	15.3%
Asian	1.39%	4.96%
Native American/Alaska Native	0.309%	0.868%
Multiracial	1.70%	1.55%

https://doi.org/10.1371/journal.pone.0205264.t001

A multi-parameterized artificial neural network (ANN)

Andoni *et al* have explored the theoretical limits of neural networks with two hidden layers and have shown their ability to represent polynomial functions [27]. Thus we hypothesized that a two-layered neural network with a sufficient number of inputs and neurons would be able to make accurate cancer risk.

Based on success with non-melanoma skin cancer [28] we used 12 neurons per hidden layer. However we also explored networks with various numbers of neurons (6, 10, 11, 13, and 20) in each layer, none of which preformed significantly better.

A schematic of our ANN is shown in Fig 1. Our ANN uses a backpropagation algorithm with bias terms and gradient descent (simultaneously using all examples in the training dataset each epoch) [29]. Inputs were normalized to fall in between 0 and 1 with sigmoidal activation functions being used throughout. A modification was made to allow further speedup of convergence by increasing the learning rate 1% each time the cost function decreases and decreasing the learning rate 5% while resetting the weights to the last iteration if the cost function increases, similar to the momentum approach [30]. The network weights were randomly initialized between -1 and 1, biases were initialized to 1, and the learning rate started at 10.

Listed in Table 2 are all the personal health inputs to our ANN. Some inputs were rescaled to comply with the mathematical format required in ANN while others take binary inputs.

With personal health information as the input, the output of our ANN was a fractional number between 0 and 1, with higher values meaning higher cancer risk. To generate the binary cancer status (Yes or No), as shown in Fig 1, it is standard practice to use a cutoff of 0.5, above which our ANN predicts a Yes cancer status. However, the much larger number of non-cancer cases in our data biases the output towards 0. So instead, once the training was complete, we than calculated sensitivity and specificity for the full range of possible cutoff values. Using the training set we selected the cutoff that maximized the sum of sensitivity and specificity. That same value was then applied to the validation data.

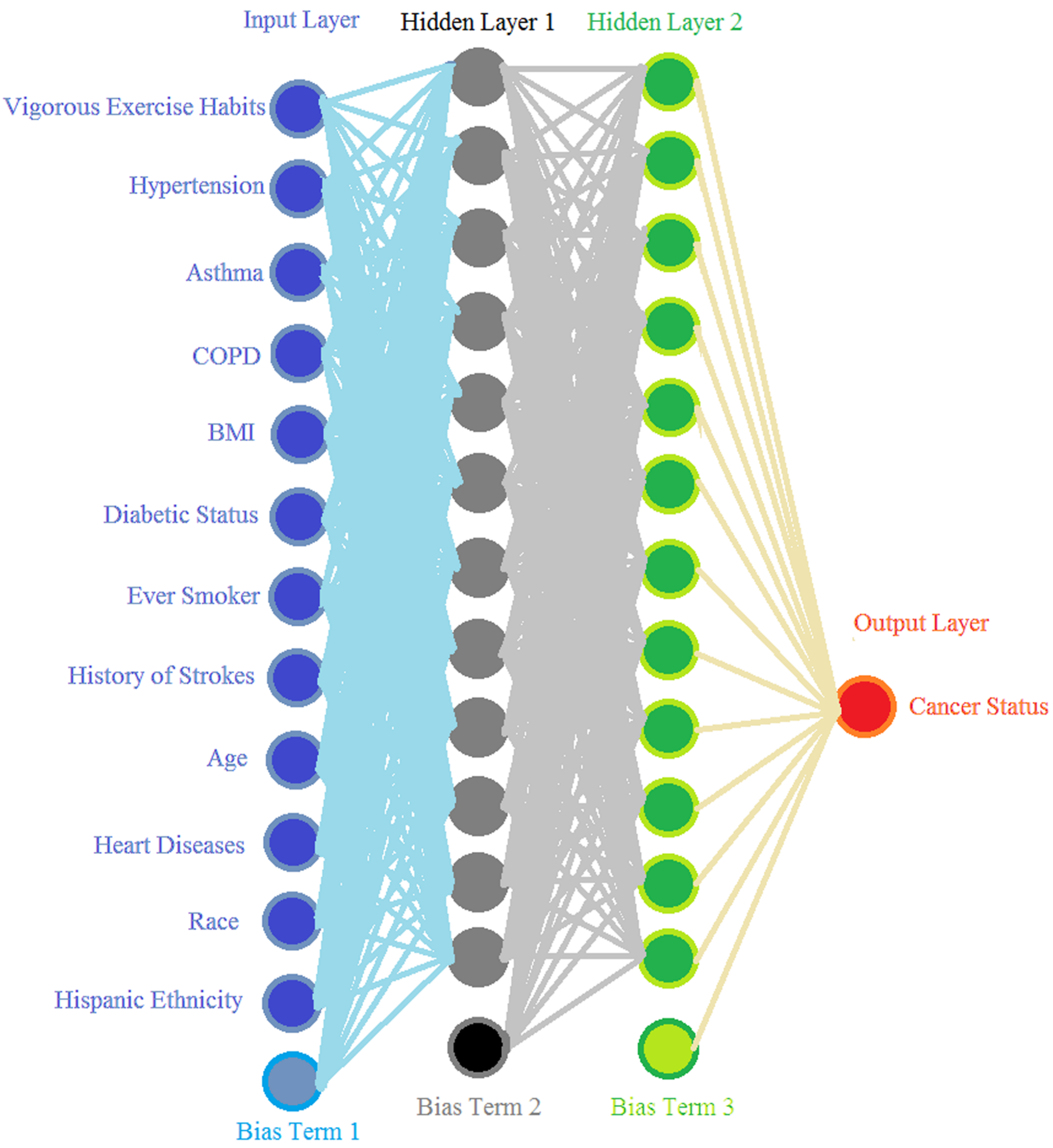

Fig 1. A sketch of our ANN. All lines are weights connecting one layer to next, with each circle either being an input, neuron, or output. The bias terms are analogous to intercepts and they improve the model's performance.

https://doi.org/10.1371/journal.pone.0205264.g001

We have also applied random forest (RF) and support vector machine (SVM) algorithms to this dataset and compared them to our ANN.

Results

Sensitivity, specificity, and AUC of the neural network

For the training set, the sensitivity was 79.8% (95% CI [31], 75.9%-83.6%) and the specificity was 79.9% (95% CI [31], 79.8%-80.1%). The validation set had sensitivity of 75.3% (95% CI [31], 68.9%-81.6%) and specificity of 80.6% (95% CI [31], 80.3%-80.8%).

Table 2. A description of the inputs used in our ANN.

Input	Input Type	Input Range	Details
Age	Continuous	0-1	18-85, (85+ recorded as 85)
BMI	Continuous	0-1	BMI of 99.95+ recorded as 99.95
Heart Disease Score	Continuous	0-1	Coronary heart disease, Angina, Heart attacks, and other heart complications each contribute 0.25 to the score
Vigorous Exercise	Continuous	0-1	Number of times per week vigorous exercise is performed; 28+ is treated as 28. Minimum time for exercise to count was 10 minutes, except for the first half of 1997 for which it was 20 minutes.
Gender	Binary	0 or 1	0 is a man and 1 is a woman
Ever Smoked	Binary	0 or 1	Never smoked is 0 and current and former smokers are 1
Emphysema	Binary	0 or 1	No COPD is 0 and COPD is 1
Asthma	Binary	0 or 1	No asthma is 0 and asthma is 1
Diabetes	Binary	0 or 1	Non-diabetics and pre-diabetics are 0, with diabetics being 1
Strokes	Binary	0 or 1	No stroke is 0 and a prior stroke is 1
Hypertension	Binary	0 or 1	No hypertension is 0, and having single measurement of it is 1
Hispanic Ethnicity	Binary	0 or 1	Non-Hispanic is 0 and Hispanic is 1
Race	Continuous	0-1	Each race is assigned a value equal to its fractional percentage in the sample plus the fractional percentage of each less common race being added to the race of interest

https://doi.org/10.1371/journal.pone.0205264.t002

Since the program computes both sensitivity and specificity for both the training and validations sets, it is important to show how they vary as function of the cutoff value. These results are shown in Fig 2.

This information is also conveyed though a conventional receiver operating characteristic (ROC) plot for both the training and validation sets in Fig 3. Our training and validation sets yielded AUC values of 0.86 (95% CI 0.85-0.88) and 0.86 (95% CI 0.84-0.89), respectively.

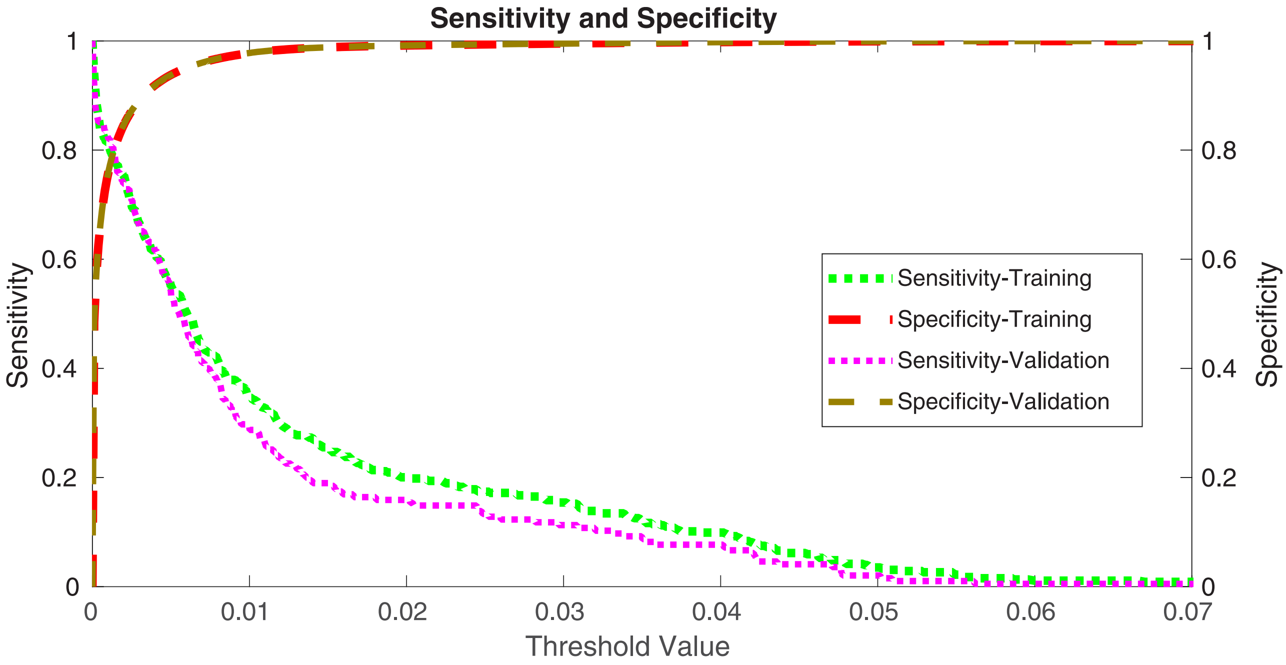

Fig 2. The sensitivity and specificity for the training and validation datasets as functions of the cutoff values.

https://doi.org/10.1371/journal.pone.0205264.g002

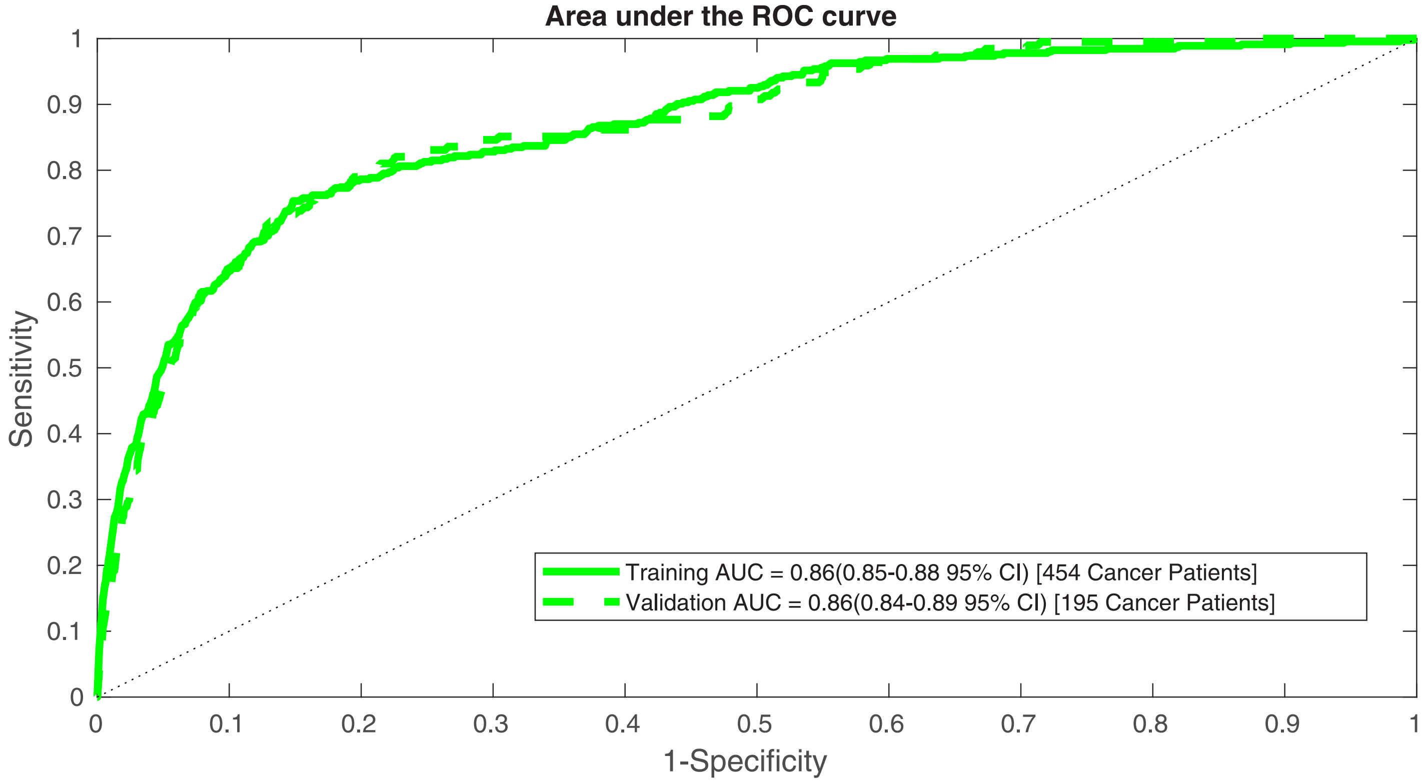

Fig 3. An ROC plot for our ANN's training and validation datasets.

https://doi.org/10.1371/journal.pone.0205264.g003

We also applyed random forest and support vector machine to this same dataset. Both of these methods did better on the training data, with RF having an AUC of 1.00 (95% CI 1.00-1.00) and SVM's AUC being 0.96 (95% CI 0.95-0.97). However neither of these methods generalized well. The performance of SVM on the validation dataset yielded an AUC 0.55 (95% CI 0.51-0.58). The performance of the RF was better with an AUC of 0.81 (95% CI 0.78-0.84) approaching the performance of the ANN. However the ANN performed the best on the validation dataset and generalized the best (Fig 4).

A risk stratification tool

While the above results indicate that our model could work well has a diagnostic test, our stated goal was to stratify cancer risk in order to improve screening selection. Accordingly, we present a simple example of how our ANN can be used to do so. When running the ANN instead of applying a cutoff and getting a binary answer, we keep the continuous output from our model and normalize it based on the maximum output from the training and validation sets. This transforms the model output to a percentage equivalent to the cancer risk.

As shown in Fig 5, we select two risk boundaries that break the cancer risk into 3 categories: high risk (represented by red), medium risk (yellow), and low risk (green). In this scheme high risk people should be screened immediately, while medium risk people should receive their standard regular screenings (per the ACS recommendations), and low risk people could be screened less frequently. We chose the boundary between medium and high risk so that only 1% of the individuals without cancer would be classified as high risk. Likewise, the boundary between low and medium risk was chosen such that only 1% of the individuals with cancer would be classified as low risk.

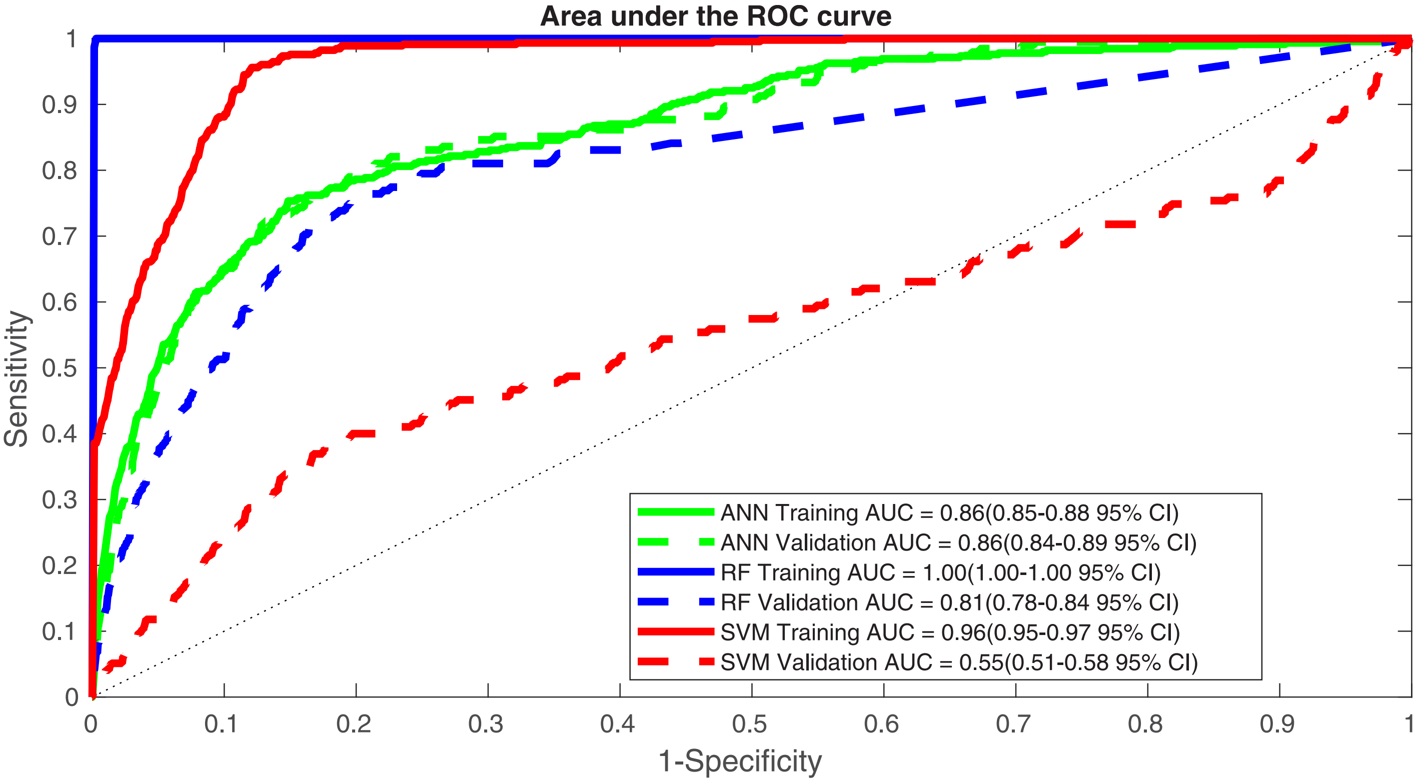

Fig 4. An ROC plot for our ANN's training and validation datasets as well as the performance of Random Forest and Support Vector Machine.

https://doi.org/10.1371/journal.pone.0205264.g004

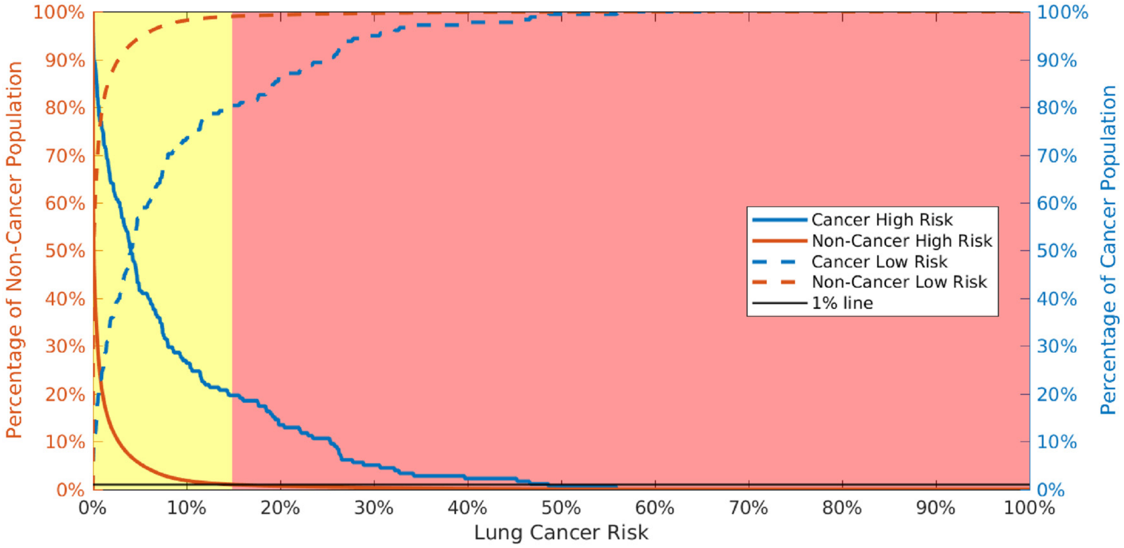

Fig 5. Cumulative distribution function for high risk (solid line) and low risk (dashed line) population without cancer (orange) and population with cancer (blue) populations in the validation dataset. Allowing for a 1% misclassification rate (black line), we can divide individual cancer risk into 3 categories: high (red), medium (yellow), and low (green, too narrow to see on the left of this figure).

https://doi.org/10.1371/journal.pone.0205264.g005

Table 3. NHIS 2016 data risk stratification results by our ANN.

	# People	# Low Risk	% Low Risk	# Medium Risk	% Medium Risk	# High Risk	% High Risk
Cancer	55	1	1.82%	44	80.0%	10	18.2%
Non-Cancer	27,844	3,362	12.1%	24,159	86.8%	323	1.16%

https://doi.org/10.1371/journal.pone.0205264.t003

We tested the risk stratification scheme with the 2016 NHIS data (27,899 individuals) which was not included in the training or validation of the model. This scheme classifies 18.2% of the population with cancer as high risk who should be screened as soon as possible and 80.0% as medium risk who should be screened according to ACS guidelines. For the population without cancer population, this scheme classifies 12.1% of the population as low risk and 86.8% medium risk (see Table 3). Effectively, our model can be used as a risk stratification tool for clinical decision support.

Discussion

The USPSTF recommends screening with LDCT for early detection of lung cancer in high risk individuals aged 55—80 who have a smoking history of at least 30 pack years and who are current smokers or have quit within the last 15 years [4]. These recommendations are endorsed by the American Cancer Society [6]. The effectiveness of LDCT screening in terms of reducing lung cancer-specific mortality was demonstrated in the National Lung Screening Trial as a 20% reduction in mortality compared to screening with chest radiograph [32]. This mortality benefit has not been reproduced in other studies, but likely reflects the lack of sufficient power in studies smaller than the NLST to do so [33–37]. The NLST study showed a sensitivity of 93.8%, but with a false positive rate of 96.4%. This high false positive rate leads to a significant number of follow-ups, ranging from additional imaging to more invasive procedures such as biopsies. Given the risk of false positive findings, the magnitude of follow-ups generated, and the potential for harm related to possible invasive interventions and additional radiation exposure, the American Cancer Society and other invested organizations recommend screening only for high risk patients at clinics with "access to high-volume, high-quality lung cancer screening and treatment centers", and only after a discussion relating "the potential benefits, limitations, and harms associated with screening for lung cancer with LDCT." [4, 6, 35, 38]

Recognizing the limitations posed by LDCT screening, other modalities for lung cancer screening continue to be investigated. These include other imaging modalities [33, 39–44], breath analysis [45], blood tests [46], urine analysis [47], biomarkers [48, 49], and genetic markers [50–55]. While many of these show promise in small studies, few have been tested in large trials. A detailed review of these methods is beyond the scope of this study, however we provide a brief summary of the accuracy, pros, and cons of each method with the corresponding references in Table 4.

The majority of these methods are proposed as adjuncts to CT screening, to refine the identification of high risk individuals who would most benefit and exclude individuals who would not. Improving identification of appropriate at-risk populations would lower unnecessary radiation exposure and expense as well as relieve the burden of the follow-up examinations, stress, and anxiety from the many false positives of LDCT screening. Our results demonstrate that an artificial neural network can fulfill a similar function, but relaying on data already gathered and a standard computer it is cheap and easy to execute. It is a non-invasive method of predicting lung cancer risk using personal health information (age, BMI, smoking, heart

Table 4. The various screening methods, with their sensitivities and specificities.

Method	Sensitivity	Specificity	Pros and Cons
Our developed ANN	75.3%	80.6%	Noninvasive, Cost-effective, Easy to implement; Less Sensitive than LDCT
Low-Dose CT Scan [56]	93.8%*	73.4%*	Noninvasive, High sensitivity; Expensive, False positives, Radiation exposure
Chest X-ray [56]	73.5%*	91.3%*	Noninvasive; Expensive, False positives, Radiation exposure
Sputum Cytology [48]	16%	99.1%	Noninvasive; Low sensitivity
Automated Sputum Cytometry [48]	40%	91%	Noninvasive, High through-put; Low sensitivity
hnRNP A2/B1 Expression [50]	80.5%	73.5%	High accuracy; Expensive
Promoter Hypermethylation [52]	63%-86%	75%-92%	High accuracy; Expensive
Microarray Gene Spectorometry [45]	80%	84%	High accuracy; Expensive, More invasive
Gas Chromatography-Mass Spectorometry [45]	51%-96.5%	66.7%-100%	Noninvasive, Can be accurate; Expensive, Difficult to perform correctly
Electronic Noses [45]	71.4%-87%	48%-100%	Noninvasive, Can be accurate; Expensive, Difficult to perform correctly
Biomarkers in Blood [46]	41%-77%	80%-93%	Approaching high accuracy; Blood draw and analysis
Buccal Mucosa Analysis [44]	79%	83%	Noninvasive, Quick; Limited testing
Urine Analysis [47]	72%-79%	85%-100%	Noninvasive, High accuracy; Limited testing

* These values are based on three years of screening and follow up on a positive screen. Instead considering each scan or radiograph in isolation the false positive rate goes way up (96.4% and 94.5%) and the positive predictive value drops to 3.8% and 5.7% for Low-Dose CT scans and chest X-rays, respectively [32, 56].

https://doi.org/10.1371/journal.pone.0205264.t004

diseases, etc.), and can be distributed into the clinics to support clinical decision-making. The ANN has the advantages of using readily available data with minimal cost and is non-invasive.

Despite the limited amount of data used in our model, it performs very well in identifying patients at risk for lung cancer. With a sensitivity of 75.3% and a specificity of 80.6%, it outperforms all the non-invasive methods listed in Table 4 except the best breath analysis tests. It performs better than chest radiograph and is competitive with many of the other, more intensive methods (e.g., blood test, microarray gene expression, hnRNP A2/B1 Expression). While the ANN does not have as high of a sensitivity as LDCT screening in the USPSTF-selected population, its low false positive rate gives a positive predictive value 10-fold higher than the CT scan (0.395 vs. 0.038).

Furthermore, we showed that the ANN can produce a continuous risk value that can be used for stratification. We presented a simple 3-tiered system that identified almost 20% of the population who could be prioritized for screening. This scheme also classified the majority of those without cancer as medium risk, who we suggest could follow the ACS guidelines for screening. Following the ACS guidelines, based on age and number of pack-years smoked, will likely recommend many of those people not receive LDCT screening. This is equivalent to identifying them as low risk in our scheme. Ideally our model would be able to do this on its own, but our model only uses if someone has ever smoked not the amount they have smoked. We hope to improve our model with more detailed smoking habit information included in the future, but even without it our model still identifies the highest risk population well.

Conclusion

We have developed and validated a multi-parameterized artificial neural network for lung cancer risk prediction based solely on personal health information readily available in EMR systems. Our results demonstrate that our artificial neural network can offer high specificity and modest sensitivity for identification of lung cancer risk, as compared to other risk predictive modalities currently employed. This approach is cheap, non-invasive, and easy to implement. While our neural network could be potentially used as a clinical tool for lung cancer risk stratification, further improvement with more risk factors included and more clinical testing would be needed.

Acknowledgments

We would like to thank Lynn Tanoue for her careful review of our manuscript and her excellent feedback.

Author Contributions

Conceptualization: Gregory R. Hart, David A. Roffman.

Formal analysis: Gregory R. Hart, David A. Roffman.

Project administration: Jun Deng.

Resources: Jun Deng.

Supervision: Jun Deng.

Writing – original draft: Gregory R. Hart, David A. Roffman.

Writing – review & editing: Gregory R. Hart, Roy Decker, Jun Deng.

References

1. American Cancer Society. Key Statistics for Lung Cancer; 2017. Available from: https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html.
2. American Cancer Society. Lung Cancer; 2016. Available from: https://www.cancer.org/cancer/lung-cancer.html.
3. Centers for Disease Control and Prevention. What Screening Test Are There For Lung Cancer; 2017. Available from: https://www.cdc.gov/cancer/lung/basic_info/screening.htm.
4. de Koning HJ, Meza R, Plevritis SK, ten Haaf K, Munshi VN, Jeon J, et al. Benefits and harms of computed tomography lung cancer screening strategies: a comparative modeling study for the U.S. Preventive Services Task Force. Annals of internal medicine. 2014; 160(5):311–20. https://doi.org/10.7326/M13-2316 PMID: 24379002
5. Centers for Disease Control and Prevention. What Are the Risk Factors for Lung Cancer; 2017. Available from: https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm.
6. American Cancer Society. Cancer Facts and Figures 2017. American Cancer Society; 2017. Available from: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2017/cancer-facts-and-figures-2017.pdf.
7. Kovalchik SA, Tammemagi M, Berg CD, Caporaso NE, Riley TL, Korch M, et al. Targeting of Low-Dose CT Screening According to the Risk of Lung-Cancer Death. New England Journal of Medicine. 2013; 369(3):245–254. https://doi.org/10.1056/NEJMoa1301851 PMID: 23863051
8. Tammemägi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection Criteria for Lung-Cancer Screening. New England Journal of Medicine. 2013; 368(8):728–736. https://doi.org/10.1056/NEJMoa1211776 PMID: 23425165
9. Katki HA, Kovalchik SA, Berg CD, Cheung LC, Chaturvedi AK. Development and Validation of Risk Models to Select Ever-Smokers for CT Lung Cancer Screening. JAMA. 2016; 315(21):2300. https://doi.org/10.1001/jama.2016.6255 PMID: 27179989
10. Iyen-Omofoman B, Tata LJ, Baldwin DR, Smith CJ, Hubbard RB. Using socio-demographic and early clinical features in general practice to identify people with lung cancer earlier. Thorax. 2013; 68(5):451–459. https://doi.org/10.1136/thoraxjnl-2012-202348 PMID: 23321602
11. Spitz MR, Etzel CJ, Dong Q, Amos CI, Wei Q, Wu X, et al. An expanded risk prediction model for lung cancer. Cancer prevention research (Philadelphia, Pa). 2008; 1(4):250–4. https://doi.org/10.1158/1940-6207.CAPR-08-0060
12. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, et al. Variations in Lung Cancer Risk Among Smokers. JNCI: Journal of the National Cancer Institute. 2003; 95(6):470–478. https://doi.org/10.1093/jnci/95.6.470 PMID: 12644540

13. Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, et al. Lung Cancer Risk Prediction: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Models and Validation. JNCI: Journal of the National Cancer Institute. 2011; 103(13):1058–1068. https://doi.org/10.1093/jnci/djr173 PMID: 21606442

14. Meza R, Hazelton WD, Colditz GA, Moolgavkar SH. Analysis of lung cancer incidence in the nurses' health and the health professionals' follow-up studies using a multistage carcinogenesis model. Cancer Causes & Control. 2008; 19(3):317–328. https://doi.org/10.1007/s10552-007-9094-5

15. Hazelton WD, Jeon J, Meza R, Moolgavkar SH. Chapter 8: The FHCRC Lung Cancer Model. In: Risk Analysis. vol. 32. Wiley/Blackwell; 2012. p. S99–S116. Available from: http://doi.wiley.com/10.1111/j.1539-6924.2011.01681.x.

16. Gray EP, Teare MD, Stevens J, Archer R. Risk Prediction Models for Lung Cancer: A Systematic Review. Clinical Lung Cancer. 2016; 17(2):95–106. https://doi.org/10.1016/j.cllc.2015.11.007 PMID: 26712102

17. Centers for Disease Control and Prevention. NHIS—Data, Questionnaires and Related Documentation; 2017. Available from: https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm.

18. Centers for Disease Control and Prevention. NHIS—About the National Health Interview Survey; 2017. Available from: https://www.cdc.gov/nchs/nhis/about_nhis.htm.

19. Duan P, Hu C, Quan C, Yi X, Zhou W, Yuan M, et al. Body mass index and risk of lung cancer: Systematic review and dose-response meta-analysis. Scientific Reports. 2015; 5(1):16938. https://doi.org/10.1038/srep16938 PMID: 26582414

20. Zhu L, Cao H, Zhang T, Shen H, Dong W, Wang L, et al. The Effect of Diabetes Mellitus on Lung Cancer Prognosis. Medicine. 2016; 95(17):e3528. https://doi.org/10.1097/MD.0000000000003528 PMID: 27124062

21. Raviv S, Hawkins KA, DeCamp MM, Kalhan R. Lung Cancer in Chronic Obstructive Pulmonary Disease. American Journal of Respiratory and Critical Care Medicine. 2011; 183(9):1138–1146. https://doi.org/10.1164/rccm.201008-1274CI PMID: 21177883

22. García Sanz MT, González Barcala FJ, Álvarez Dobaño JM, Valdés Cuadrado L. Asthma and risk of lung cancer. Clinical and Translational Oncology. 2011; 13(10):728–730. https://doi.org/10.1007/s12094-011-0723-9 PMID: 21975334

23. Centers for Disease Control and Prevention. Lung Cancer Rates by Race and Ethnicity; 2017. Available from: https://www.cdc.gov/cancer/lung/statistics/race.htm.

24. Lindgren A, Pukkala E, Nissinen A, Tuomilehto J. Blood pressure, smoking, and the incidence of lung cancer in hypertensive men in North Karelia, Finland. American journal of epidemiology. 2003; 158 (5):442–7. https://doi.org/10.1093/aje/kwg179 PMID: 12936899

25. Moore SC, Lee IM, Weiderpass E, Campbell PT, Sampson JN, Kitahara CM, et al. Association of Leisure-Time Physical Activity With Risk of 26 Types of Cancer in 1.44 Million Adults. JAMA Internal Medicine. 2016; 176(6):816. https://doi.org/10.1001/jamainternmed.2016.1548 PMID: 27183032

26. Chen PC, Muo CH, Lee YT, Yu YH, Sung FC. Lung Cancer and Incidence of Stroke: A Population-Based Cohort Study. Stroke. 2011; 42(11):3034–3039. https://doi.org/10.1161/STROKEAHA.111.615534 PMID: 21903961

27. Andoni A, Panigrahy R, Valiant G, Zhang L. Learning Polynomials with Neural Networks. In: Proceedings of teh 31st International Conference on Machine Learning. Beijing, China: JMLR; 2014. Available from: http://proceedings.mlr.press/v32/andoni14.pdf.

28. Roffman D, Hart GR, Girardi M, Ko CJ, Deng J. Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. Scientific Reports. 2018; 8(1):1701. https://doi.org/10.1038/s41598-018-19907-9 PMID: 29374196

29. Stanford University. Multi-Layer Neural Network; 2015. Available from: http://ufldl.stanford.edu/tutorial/supervised/MultiLayerNeuralNetworks/.

30. Rumelhart DE, McClelland JL, University of California SDPRG. Learning internal representations by error propagation. In: Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. MIT Press; 1986. p. 318–362. Available from: http://dl.acm.org/citation.cfm?id=104293.

31. Epi Tools—Calculate confidence limits for a sample proportion; 2017. Available from: http://epitools.ausvet.com.au/content.php?page=CIProportion.

32. Berg CD. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. The new england journal of medicine. 2011; 365(5):395–409. https://doi.org/10.1056/NEJMoa1102873 PMID: 21714641

33. Doria-Rose VP, Szabo E. Screening and Prevention of Lung Cancer. In: Kernstine Kemp H and Reckamp Karen L, editor. Lung Cancer: A Multidisciplinary Approach to Diagnosis and Management. 1st ed.

New York: Demos Medical Publishing; 2011. p. 53–72. Available from: https://books.google.com/books?id=tq9vl31etFUC{&}pg=PA53{&}source=gbs_toc_r{&}cad=4{#}v=onepage{&}q{&}f=false.

34. Infante M, Cavuto S, Lutman FR, Brambilla G, Chiesa G, Ceresoli G, et al. A Randomized Study of Lung Cancer Screening with Spiral Computed Tomography. American Journal of Respiratory and Critical Care Medicine. 2009; 180(5):445–453. https://doi.org/10.1164/rccm.200901-0076OC PMID: 19520905

35. Bach PB, Mirkin JN, Oliver TK, Azzoli CG, Berry DA, Brawley OW, et al. Cost-effectiveness of computed tomography lung cancer screening. JAMA. 2012; 307(22):2418—2429.

36. Lopes Pegna A, Picozzi G, Mascalchi M, Maria Carozzi F, Carrozzi L, Comin C, et al. Design, recruitment and baseline results of the ITALUNG trial for lung cancer screening with low-dose CT. Lung cancer (Amsterdam, Netherlands). 2009; 64(1):34–40. https://doi.org/10.1016/j.lungcan.2008.07.003

37. Pedersen JH, Ashraf H, Dirksen A, Bach K, Hansen H, Toennesen P, et al. The Danish Randomized Lung Cancer CT Screening Trial— Overall Design and Results of the Prevalence Round. JTO Acquisition. 2009; 4:608–614.

38. Shead DA, Corrigan A, Hanisch LJ, Clarke R, Kidney S, Williams K. NCCN Guidelines For Patients: Lung Cancer Screening. Fort Washington: National Comprehensive Cancer Network; 2017. Available from: https://www.nccn.org/patients/guidelines/content/PDF/lung_screening.pdf.

39. Brett GZ. Earlier diagnosis and survival in lung cancer. British medical journal. 1969; 4(5678):260–2. https://doi.org/10.1136/bmj.4.5678.260 PMID: 5345935

40. Doria-Rose VP, Marcus PM, Szabo E, Tockman MS, Melamed MR, Prorok PC. Randomized controlled trials of the efficacy of lung cancer screening by sputum cytology revisited. Cancer. 2009; 115 (21):5007–5017. https://doi.org/10.1002/cncr.24545 PMID: 19637354

41. Kubík AK, Parkin DM, Zatloukal P. Czech Study on Lung Cancer Screening: post-trial follow-up of lung cancer deaths up to year 15 since enrollment. Cancer. 2000; 89(11 Suppl):2363–8. PMID: 11147613

42. Marcus PM, Bergstralh EJ, Zweig MH, Harris A, Offord KP, Fontana RS. Extended Lung Cancer Incidence Follow-up in the Mayo Lung Project and Overdiagnosis. JNCI Journal of the National Cancer Institute. 2006; 98(11):748–756. https://doi.org/10.1093/jnci/djj207 PMID: 16757699

43. Manser RL, Irving LB, Byrnes G, Abramson MJ, Stone CA, Campbell DA. Screening for lung cancer: a systematic review and meta-analysis of controlled trials. Thorax. 2003; 58(9):784–9. https://doi.org/10.1136/thorax.58.9.784 PMID: 12947138

44. Radosevich AJ, Mutyal NN, Rogers JD, Gould B, Hensing TA, Ray D, et al. Buccal Spectral Markers for Lung Cancer Risk Stratification. PLoS ONE. 2014; 9(10):e110157. https://doi.org/10.1371/journal.pone.0110157 PMID: 25299667

45. Nardi-Agmon I, Peled N. Exhaled breath analysis for the early detection of lung cancer: recent developments and future prospects. Lung Cancer (Auckland, NZ). 2017; 8:31–38.

46. Macdonald IK, Parsy-Kowalska CB, Chapman CJ. Autoantibodies: Opportunities for Early Cancer Detection. Trends in Cancer. 2017; 3(3):198–213. https://doi.org/10.1016/j.trecan.2017.02.003 PMID: 28718432

47. Nolen BM, Lomakin A, Marrangoni A, Velikokhatnaya L, Prosser D, Lokshin AE. Urinary protein biomarkers in the early detection of lung cancer. Cancer prevention research (Philadelphia, Pa). 2015; 8 (2):111–9. https://doi.org/10.1158/1940-6207.CAPR-14-0210

48. Kemp RA, Reinders DM, Turic B, Kemp R. Detection of Lung Cancer by Automated Sputum Cytometry. Journal of Thoracic Oncology. 2007; 2:993–1000. https://doi.org/10.1097/JTO.0b013e318158d488 PMID: 17975489

49. Palcic B, Garner DM, Beveridge J, Sun XR, Doudkine A, MacAulay C, et al. Increase of sensitivity of sputum cytology using high-resolution image cytometry: Field study results. Cytometry. 2002; 50 (3):168–176. https://doi.org/10.1002/cyto.10065 PMID: 12116340

50. Tockman MS, Mulshine JL, Piantadosi S, Erozan YS, Gupta PK, Ruckdeschel JC, et al. Prospective detection of preclinical lung cancer: results from two studies of heterogeneous nuclear ribonucleoprotein A2/B1 overexpression. Clinical Cancer Research. 1997; 3(12).

51. Belinsky SA, Liechty KC, Gentry FD, Wolf HJ, Rogers J, Vu K, et al. Promoter Hypermethylation of Multiple Genes in Sputum Precedes Lung Cancer Incidence in a High-Risk Cohort. Cancer Research. 2006; 66(6):3338–3344. https://doi.org/10.1158/0008-5472.CAN-05-3408 PMID: 16540689

52. Hulbert A, Jusue-Torres I, Stark A, Chen C, Rodgers K, Lee B, et al. Early Detection of Lung Cancer Using DNA Promoter Hypermethylation in Plasma and Sputum. Clinical Cancer Research. 2017; 23(8). https://doi.org/10.1158/1078-0432.CCR-16-1371 PMID: 27729459

53. Beau-Faller M, Legrain M, Voegeli AC, Guérin E, Lavaux T, Ruppert AM, et al. Detection of K-Ras mutations in tumour samples of patients with non-small cell lung cancer using PNA-mediated PCR clamping. British Journal of Cancer. 2009; 100(6):985–992. https://doi.org/10.1038/sj.bjc.6604925 PMID: 19293811

54. Mao L, Hruban RH, Boyle JO, Tockman M, Sidransky D. Detection of oncogene mutations in sputum precedes diagnosis of lung cancer. Cancer research. 1994; 54(7):1634–7. PMID: 8137272

55. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nature Medicine. 2007; 13(3):361–366. https://doi.org/10.1038/nm1556 PMID: 17334370

56. Church TR, Black William C, Aberle Denise R, Berg Christine D, Clingan Kathy L, Duan Fenghai, et al. Results of Initial Low-Dose Computed Tomographic Screening for Lung Cancer. N Engl J Med. 2013; 21368(23):1980–91.