Original Paper

Identifying Lung Cancer Risk Factors in the Elderly Using Deep Neural Networks: Quantitative Analysis of Web-Based Survey Data

Songjing Chen, PhD; Sizhu Wu, PhD

Institute of Medical Information and Library, Chinese Academy of Medical Sciences / Peking Union Medical College, Beijing, China

Corresponding Author:
Songjing Chen, PhD
Institute of Medical Information and Library
Chinese Academy of Medical Sciences / Peking Union Medical College
No 3, Yabao Road, Chaoyang District
Beijing
China
Phone: 86 01052328761
Email: chen.songjing@imicams.ac.cn

Abstract

Background: Lung cancer is one of the most dangerous malignant tumors, with the fastest-growing morbidity and mortality, especially in the elderly. With a rapid growth of the elderly population in recent years, lung cancer prevention and control are increasingly of fundamental importance, but are complicated by the fact that the pathogenesis of lung cancer is a complex process involving a variety of risk factors.

Objective: This study aimed at identifying key risk factors of lung cancer incidence in the elderly and quantitatively analyzing these risk factors' degree of influence using a deep learning method.

Methods: Based on Web-based survey data, we integrated multidisciplinary risk factors, including behavioral risk factors, disease history factors, environmental factors, and demographic factors, and then preprocessed these integrated data. We trained deep neural network models in a stratified elderly population. We then extracted risk factors of lung cancer in the elderly and conducted quantitative analyses of the degree of influence using the deep neural network models.

Results: The proposed model quantitatively identified risk factors based on 235,673 adults. The proposed deep neural network models of 4 groups (age ≥65 years, women ≥65 years old, men ≥65 years old, and the whole population) achieved good performance in identifying lung cancer risk factors, with accuracy ranging from 0.927 (95% CI 0.223-0.525; *P*=.002) to 0.962 (95% CI 0.530-0.751; *P*=.002) and the area under curve ranging from 0.913 (95% CI 0.564-0.803) to 0.931(95% CI 0.499-0.593). Smoking frequency was the leading risk factor for lung cancer in men 65 years and older. Time since quitting and smoking at least 100 cigarettes in their lifetime were the main risk factors for lung cancer in women 65 years and older. Men 65 years and older had the highest lung cancer incidence among the stratified groups, particularly non–small cell lung cancer incidence. Lung cancer incidence decreased more obviously in men than in women with smoking rate decline.

Conclusions: This study demonstrated a quantitative method to identify risk factors of lung cancer in the elderly. The proposed models provided intervention indicators to prevent lung cancer, especially in older men. This approach might be used as a risk factor identification tool to apply in other cancers and help physicians make decisions on cancer prevention.

(J Med Internet Res 2020;22(3):e17695) doi: 10.2196/17695

KEYWORDS

deep learning; lung cancer; risk factors; aged; primary prevention

Introduction

Background

Lung cancer is one of the most dangerous malignant tumors, with the fastest-growing morbidity and mortality, especially in the elderly. With the rapid growth of the elderly population in recent years, lung cancer prevention and control are becoming much more important than ever before. Non–small cell lung cancer (NSCLC) is the most common type of lung cancer [1].

Lung cancer pathogenesis is a complex process involving various risk factors. Factors such as smoking [2,3], secondhand smoke [4], high levels of air pollution exposure [5], and drinking water that has a high level of arsenic [6,7] can increase the risk of occurrence of lung cancer. The relationship between these risk factors and lung cancer incidence is an urgent research problem.

In high-income countries, a combination of early diagnosis, screening, and treatment has been effective in increasing population-based survival for certain cancers [8-10]. Many lung cancer screening-related studies have been conducted recently. In the United States, the National Lung Screening Trial was conducted to investigate the possibility that low-dose computed tomography (CT) could reduce lung cancer mortality [11]. Zahnd and Eberth found that use of CT screening was higher than in earlier estimates using 2017 Behavioral Risk Factor Surveillance System (BRFSS) survey data [12]. The US Preventive Services Task Force recommended annual screening of individuals at high risk of lung cancer aged 55 to 80 years who have a 30–pack-year smoking history and currently smoke or had quit within the past 15 years [13]. Berkowitz and colleagues used 2012 BRFSS data to develop multilevel small-area estimate mixed models to generate county-level estimates for 6 smoking status categories (current, some days, every day, former, ever, and never) [14].

Machine learning algorithms are being used more widely for lung cancer screening, detection, diagnosis, and other related research. Luna and colleagues used random forest as an accurate machine learning method to identify known and new predictors of symptomatic radiation pneumonitis, which is a radiotherapy dose-limiting toxicity for locally advanced NSCLC [15]. Palani and Venkatalakshmi used a fuzzy clustering method to predict lung cancer through continuous monitoring using a new internet of things and to improve health care by providing medical instructions [16]. A K-means clustering algorithm, based initially on 400 cancer and non–cancer patients' data, was developed to identify relevant and nonrelevant lung cancer data for early detection of lung cancer [17]. Liu and colleagues used multivariable logistic regression to assess the relationship between body mass index and respiratory conditions, asthma, and chronic obstructive pulmonary disease (COPD) based on BRFSS data [18]. A series of machine learning methods were applied to classify lung cancer patients' survival, including linear regression, decision trees, gradient boosting machines, support vector machines, and a custom ensemble [19]. Deep learning methods were previously rarely used to identify lung cancer risk factors, but their use has become more common recently. Cha and colleagues studied a deep convolutional neural network model to detect operable lung cancer with chest radiographs [20]. Deep learning algorithms might aid fully automated lung cancer detection even at very low effective radiation doses of 0.11 mSv [21]. Hosny and colleagues provided evidence that a convolutional neural network might be used for mortality risk stratification based on standard-of-care CT images from NSCLC patients [22].

Objective

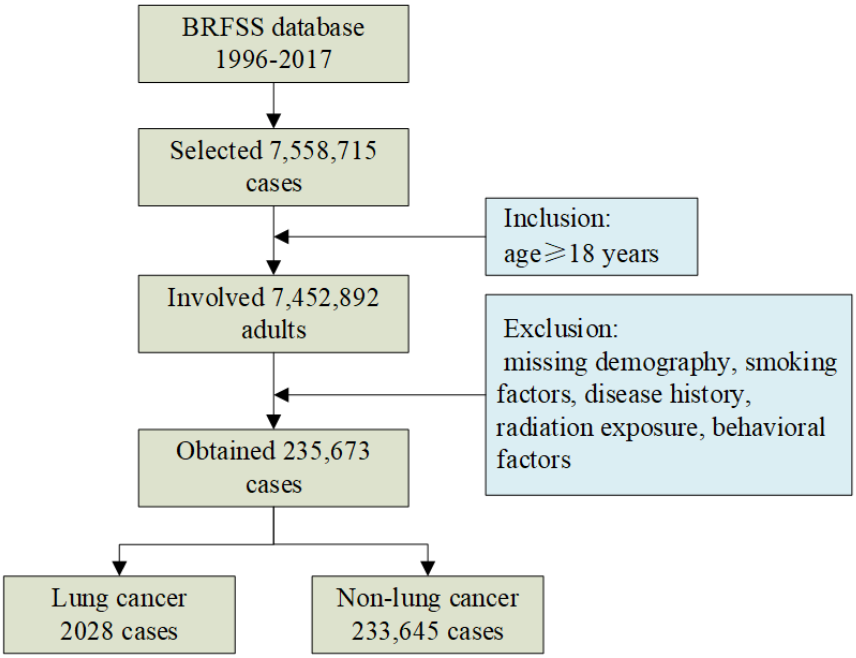
This study aimed at identifying key risk factors of lung cancer incidence in a stratified elderly population and quantitatively analyzing the risk factors' degree of influence using a deep neural network (DNN) method. Using Web-based survey data, we focused on multidisciplinary risk factors, such as smoking habit, disease history, radiation exposure, behavioral risk, environmental risk, and medical demographics. Our main research problems were how to find the leading causative factors of lung cancer incidence from complex related risk factors and to quantitatively analyze their degree of influence. Our results could help physicians in preventing lung cancer and taking effective measures for early detection.

Methods

Data Source

We obtained lung cancer risk factors from the BRFSS [23], an open access source from the US Centers for Disease Control and Prevention. BRFSS collects survey data from US residents about their health-related risk behaviors, chronic health conditions, use of preventive services, and so on. In this study, we used lung cancer behavioral health risk data of 235,673 adults from all 50 US states between 1996 and 2017. The flowchart in Figure 1 shows the data selection process.

Figure 1. Data selection flowchart. BRFSS: Behavioral Risk Factor Surveillance System.



Lung cancer has many causative factors, including age 65 years and older, body mass index, education, smoking habit, personal history of cancer, family history of cancer, CT or computerized axial tomography (CAT) scan, asthma history, and COPD history. Table 1 lists some relevant survey questions from the BRFSS questionnaire that we used to collect data for this study.

Table 1. Lung cancer risk factors assessed by the Behavioral Risk Factor Surveillance System questionnaire.

Risk factors	Description
Age	Age ≥65 years? (yes/no)
Body mass index	Level 1: <18.5 kg/m^2; 2: 18.5-24.9 kg/m^2; 3: 25.0-29.9 kg/m^2; 4: ≥30.0 kg/m^2
Education	Level of education completed (level 1: Did not graduate from high school; 2: Graduated from high school; 3: Attended postsecondary or technical school; 4: Graduated from postsecondary or technical school)
Smoked at least 100 cigarettes	Smoked at least 100 cigarettes in your entire life (yes/no; 1 pack contains 20 cigarettes)
Smoking frequency	Level 1: Every day; 2: Some days; 3: Not at all
Smoking start age	How old were you when you first started to smoke cigarettes regularly? (Age in years)
Smoking intensity	How many cigarettes do you smoke each day? (Number of cigarettes/day)
Smoking quit attempts	During the past 12 months, have you stopped smoking for 1 day or longer? (yes/no)
Time since quitting	How long has it been since you last smoked a cigarette? (1: Within the past month; 2: Within the past 3 months; 3: Within the past 6 months; 4: Within the past year; 5: Within the past 5 years; 6: Within the past 10 years; 7: 10 years or more; 8: Never smoked regularly)
E-cigarette use	Have you ever used an e-cigarette or other electronic vaping product, even just one time? (yes/no)
E-cigarette use frequency	Do you now use e-cigarettes or other electronic vaping products every day, some days, or not at all? (1: Every day; 2: Some days; 3: Not at all)
Chronic obstructive pulmonary disease (COPD) history	History of COPD (yes/no)
Asthma history	History of asthma (yes/no)
Cancer history	Personal history of cancer (yes/no)
Family history of cancer	Family history of cancer (yes/no)
Computed tomography (CT) or computerized axial tomography (CAT) scan	In the last 12 months, did you have a CT or CAT scan? (yes/no)

Participants who were 65 years and older accounted for about 35.01% (82,503/235,673) of the survey population and those aged 18 to 64 years accounted for 64.99% (153,170/235,673). By sex, 53.99% (127,262/235,673) were women and 46.01% (108,411/235,673) were men.

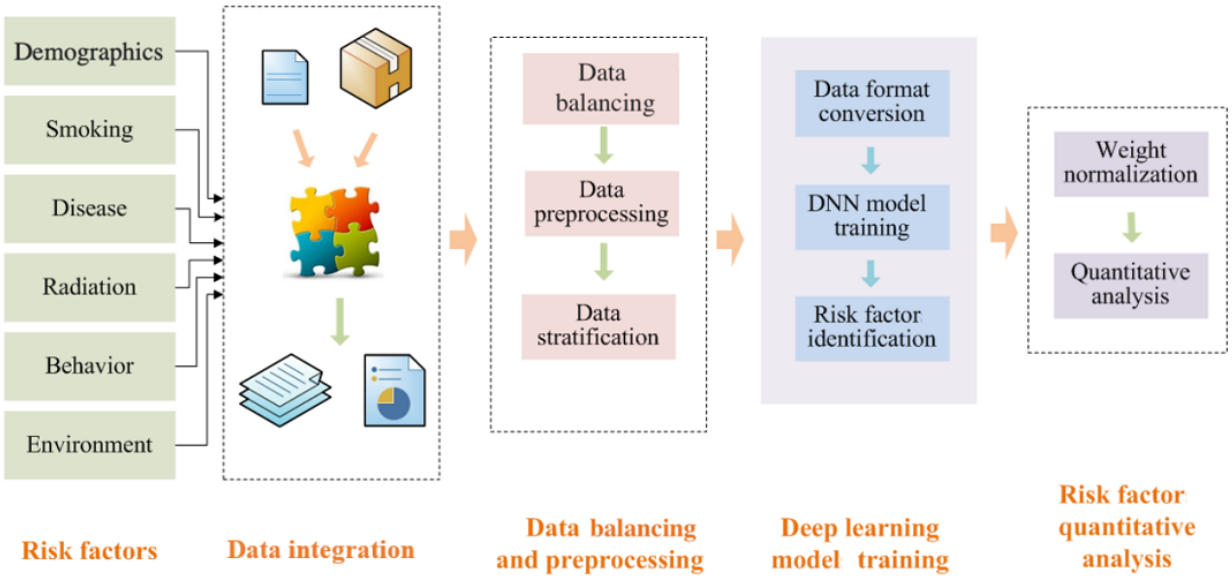
We derived environmental risk factors from the open access website of the US Environmental Protection Agency [24], including air pollutants and drinking water. According to the investigation date, we linked environmental data with risk factors from the BRFSS.

Data Analysis

Overview

In this study, we employed a DNN model to identify risk factors for lung cancer in the elderly. The DNN model had a multiple-layer structure and powerful data expression ability. In particular, in training models based on the large dataset, DNN had high calculation accuracy. First, we integrated the data on medical demographics, smoking habit, disease history, radiation exposure, behavioral risk, and other aspects. Second, since the number of cases of lung cancer was much lower than that of non–lung cancers, we balanced the data. Then we preprocessed these balanced data. Third, we trained DNN models by leveraging the stratified data of the elderly population. We extracted the stratified risk factors through DNN models. Fourth, we developed a quantitative analysis of the degree of effect of the risk factors in elderly patients. Therefore, the whole study comprised 4 sections: data integration, data balancing and preprocessing, training of DNN models, and quantitative analysis of risk factors, as Figure 2 shows.

Figure 2. Schematic diagram of lung cancer risk factor identification in the elderly. DNN: deep neural network.



Data Integration

Lung cancer incidence is caused by multiple risk factors [25-27], particularly in the elderly [28]. We integrated these risk factors, including medical demographics, smoking, disease history, radiation exposure, behavioral risk, and environmental risk. Medical demographic factors were age, sex, body mass index, and education level. Smoking factors were smoking intensity, age when starting to smoke, smoking frequency, time since quitting, e-cigarette use, secondhand smoke exposure, and other smoking habits. Disease history referred to COPD history, asthma history, personal cancer history, and family history of cancer. Radiation exposure involved radiotherapy of the breast or chest, CT or CAT medical imaging examination, and occupational exposure to asbestos, radon, and arsenic. We also took into account dietary and exercise habits and other behavioral risk factors.

Data Balancing and Preprocessing

The ratio of lung cancer to non–lung cancer cases was about 1:115. When studying the pathogenesis of lung cancer, this situation could cause a data imbalance problem. Therefore, we used the synthetic minority oversampling technique (SMOTE) [29] to solve the imbalance problem. SMOTE is based on the K-nearest neighbor algorithm to simulate the minority sample. We then added these simulated samples to the whole dataset.

At the same time, the integrated data had vacancy value, incompleteness, and other problems. We therefore preprocessed the data using techniques such as vacancy value filling and noise data smoothing. We used multiple imputation [30] to fill in missing values. We conducted singular value decomposition [31] to reduce data noise in the data preprocessing stage.

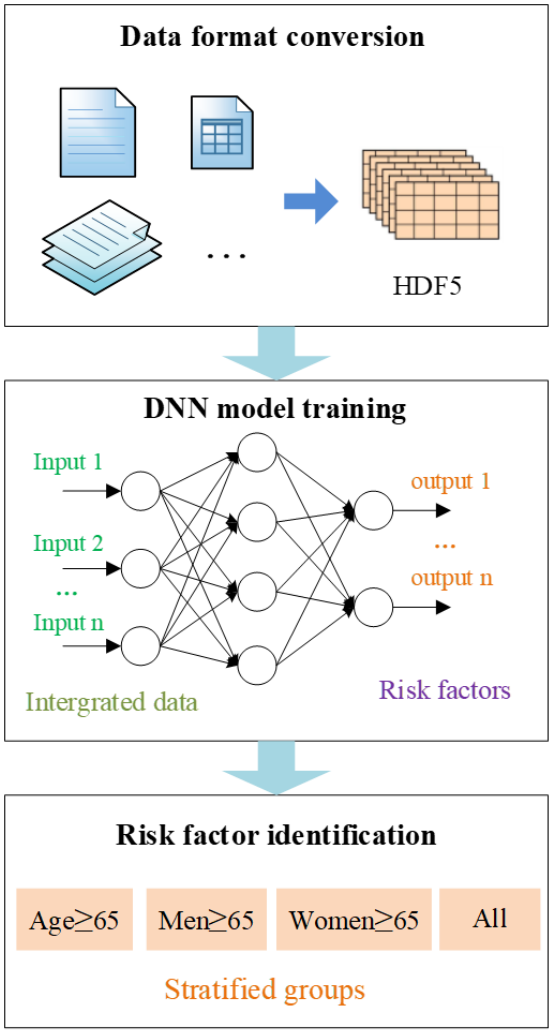
We divided the preprocessed data into 4 groups: those aged 65 years and older (age ≥65 years), women aged 65 years and older (women ≥65 years), men aged 65 years and older (men ≥65 years), and the whole population.

Deep Neural Network Model Training

By leveraging the weights of the DNN models, we quantified the degree of influence of risk factors on lung cancer incidence in the elderly (Figure 3). First, we converted the data format into hierarchical data format version 5 (HDF5) [32] in the 4

stratified groups (age ≥65 years, women ≥65 years, men ≥65 years, and the whole group) separately. HDF5 is recognized by Convolutional Architecture for Fast Feature Embedding (Caffe) [33], an open source general deep learning framework. Second, we used the Caffe framework to train DNN models based on the stratified groups in sequence. We input integrated data through an input layer, and then computed the weight values of different risk factors in a hidden layer. We obtained key risk factors using weight values through the output layer of the DNN model. Third, we extracted different risk factors of the stratified groups according to their DNN models.

Figure 3. Deep learning model training process. DNN: deep neural network; HDF5: hierarchical data format version 5.



The DNN model of the group aged 65 years and older consisted of 3 layers: the input layer, hidden layer, and output layer. This model included 1 input layer, 3 hidden layers, and 1 output layer. Layer-to-layer was fully connected. In other words, any neuron in the *i*th layer must be connected to any neuron in the *(i+1)*th layer. Therefore, there was a linear relationship where $z = \sum w_i x_i + b$, plus an activation function, $\sigma(z)$. We used a rectified linear unit function, given in Equation 1 (Figure 4), as an activation function to improve model expression ability. We used 10-fold cross-validation to test algorithm accuracy. We divided the data of the group aged 65 years and older into 10 parts. We rotated them to use 9 of them as a training dataset and 1 as a test dataset for DNN model training.

Figure 4. Data analysis equations.

$$f(x) = \begin{cases} 0, & x \le 0 \\ x, & x > 0 \end{cases} \quad (1)$$

$$\alpha^l = \sigma(\mathbf{W}^l \alpha^{l-1} + \mathbf{b}^l) \quad (2)$$

$$w_{Normalized} = \frac{w - w_{Min}}{w_{Max} - w_{Min}} \quad (3)$$

The output results α^L of the DNN model can be deduced from Equation 2 (Figure 4), where W is the weight matrix between the hidden layer and the output layer, which represents the influence of risk factors on lung cancer incidence; L is the number of layers and variable *l* is 2 to L; and b is the bias vector. The numbers of input nodes and output nodes relied on the number of input and output factors, and the hidden-layer number was determined by data size. We set a value of 30 for the input nodes, 3 for the hidden layers, and 9 for the output nodes. In this way, we constructed the DNN model of the group aged 65 years and older. We used the same network structure to train the DNN models of the other 3 stratified groups separately.

Risk Factor Quantitative Analysis

We normalized the weight (*w*) using Equation 3 (Figure 4) to extract key risk factors of lung cancer occurrence. The value of normalized weight ($w_{Normalized}$) was between 0 and 1. w_{Min} is the minimum value of weight, and w_{Max} is the maximum value of weight. We developed a quantitative analysis of different risk factors in the 4 groups. Because weights represented the degree of influence of risk factors on lung cancer occurrence, we compared the weights of risk factors to identify targeted factors among the 4 stratified groups.

Results

Risk Factor Weights

Figure 5 shows the weights of risk factors in the 4 stratified groups obtained using DNN models. Though leveraging weights of DNN models, we quantitatively analyzed the degree of the risk factors' influence on lung cancer incidence in the elderly. Table 2 shows the values of weights and odds ratios (95% CI) of these main risk factors.

Figure 5. Normalized weights of risk factors in the stratified groups. BMI: body mass index; CAT: computerized axial tomography; COPD: chronic obstructive pulmonary disease; CT: computed tomography; PM2.5: fine particulate matter with a diameter ≤2.5 μm.

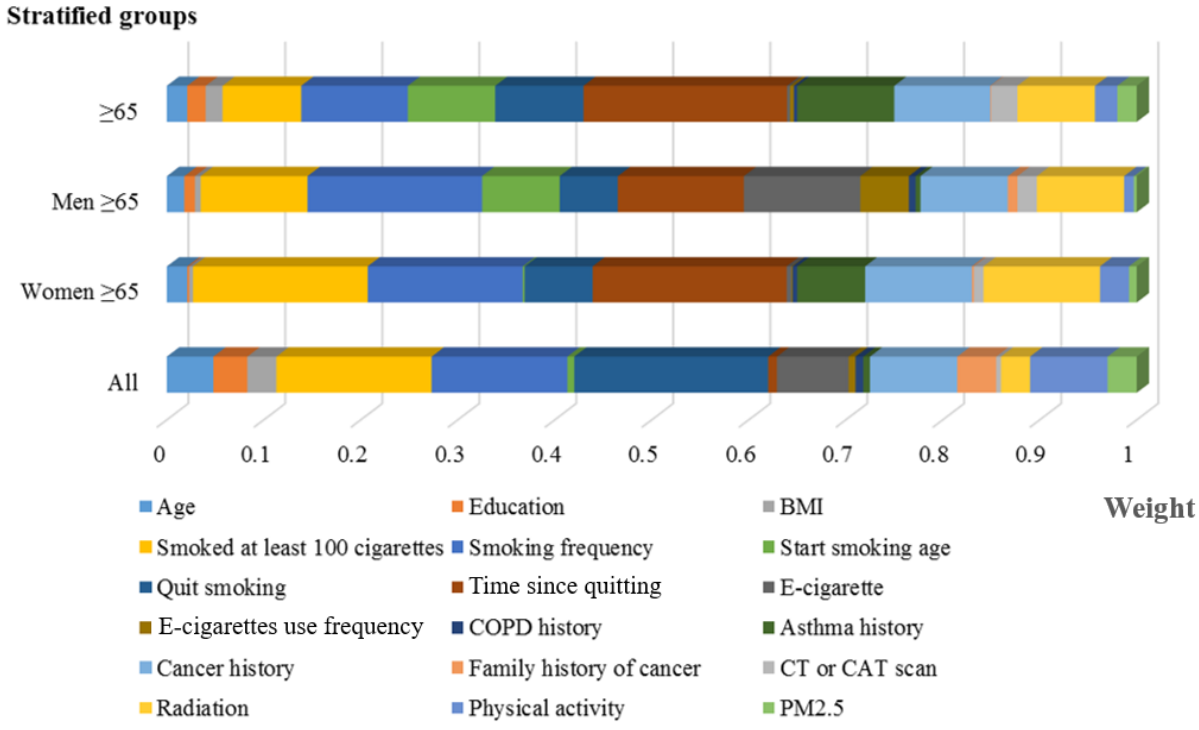


Table 2. Normalized weight values and odds ratios (95% CI) of the main risk factors in the 4 population groups.

Risk factors	Population aged ≥65 years		Men aged ≥65 years		Women aged ≥65 years		All age groups	
	Weight	Odds ratio (95% CI)	Weight	Odds ratio (95% CI)	Weight	Odds ratio (95% CI)	Weight	Odds ratio (95% CI)
Time since quitting	0.21	1.422 (0.806-1.095)	0.13	1.587 (0.776-0.998)	0.20	1.590 (0.927-1.358)	0.009	1.109 (0.993-1.322)
Smoking frequency	0.11	1.312 (0.796-0.998)	0.18	1.625 (0.866-1.097)	0.16	1.536 (1.106-1.427)	0.14	1.370 (1.352-1.701)
Cancer history	0.099	1.295 (0.876-1.027)	0.09	1.387 (1.239-1.667)	0.11	1.442 (0.951-1.356)	0.09	1.271 (0.852-1.201)
Smoking quit attempts	0.091	1.253 (0.933-1.201)	0.06	1.273 (1.413-1.702)	0.07	1.368 (1.127-1.406)	0.20	1.405 (0.995-1.381)
Lifetime smoking of ≤100 cigarettes	0.081	1.239 (1.336-1.587)	0.11	1.506 (0.681-0.937)	0.18	1.588 (1.237-1.601)	0.16	1.387 (1.225-1.611)
Asthma history	0.08	1.303 (1.029-1.403)	0.005	1.095 (0.962-1.329)	0.07	1.381 (0.953-1.317)	0.007	1.112 (0.961-1.406)
Radiation	0.08	1.224 (1.550-1.781)	0.09	1.291 (0.983-1.307)	0.12	1.453 (1.302-1.759)	0.03	1.190 (0.952-1.357)
E-cigarette use	0.023	1.025 (0.766-0.934)	0.12	1.539 (1.112-1.406)	0.005	1.135 (0.897-1.309)	0.074	1.239 (0.851-1.307)
Physical activity	0.023	1.132 (0.983-1.246)	0.01	1.170 (0.851-1.209)	0.03	1.280 (0.991-1.308)	0.08	1.268 (1.131-1.670)

Effect of Risk Factors on Lung Cancer

Those aged 65 years and older were more sensitive to how long ago former smokers had quit and smoking frequency, which were related to smoking. This correlation was more obvious in men aged 65 years and older. Those aged 65 years and older who had quit smoking for a short time or smoked more every day were prone to lung cancer.

Smoking frequency was the leading risk factor for lung cancer in men aged 65 years and older. As Table 2 shows, the weights of smoking frequency and time since quitting were 0.18 and 0.13, respectively, in this group of men. The weight of smoking frequency was 38.5% higher than the weight of time since quitting. The top 4 risk factors of men aged 65 years and older (smoking frequency, time since quitting, use of e-cigarettes, and having smoked at least 100 cigarettes in their lifetime) were all associated with smoking. These smoking-related risk factors had a greater influence than other risk factors on men who were 65 years and older. Men in this age group who actively quit smoking were more likely to avoid lung cancer.

Time since quitting and smoking at least 100 cigarettes over their lifetime were the main risk factors for lung cancer occurrence in women aged 65 years and older. As Table 2 shows, the weight of time since quitting was 0.20 in this group of women, which was 11.1% greater than the weight of having smoked at least 100 cigarettes (0.18). The top 3 relevant risk factors were associated with smoking habit factors in women aged 65 years and older: time since quitting, having smoking at least 100 cigarettes, and smoking frequency. Therefore, smoking-related risk factors had a greater influence than other risk factors on women in this age group.

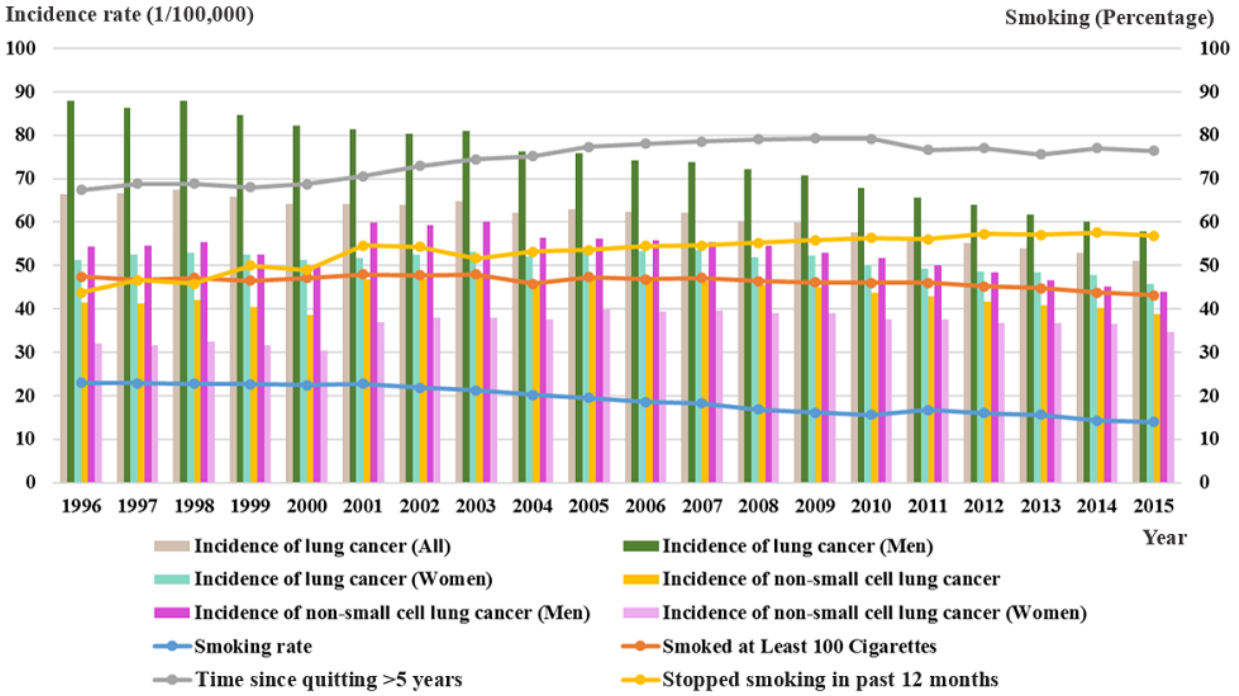
Cancer history ranked in the top risk factors in the 4 stratified groups, which may suggest that cancer history played an important role in the incidence of lung cancer [34,35]. Women aged 65 years and older were more sensitive to radiation exposure than were other groups. Physical activity was the fifth risk factor in the whole group.

Association Between Smoking and Lung Cancer Incidence

Men aged 65 years and older had the highest lung cancer incidence in these stratified groups, especially the incidence of NSCLC. We compared the incidence rate of lung cancer, NSCLC, and small cell lung cancer among all ages, under 65 years, and 65 years and older. NSCLC incidence in men 65 years and older was 286 cases per 100,000 people between 2011 and 2015, which was higher than that of women aged 65 years and older (203 per 100,000). Therefore, controlling smoking in men age 65 years and older could be more effective in preventing lung cancer.

Lung cancer incidence decreased much more rapidly in men than in women with a decline in smoking rate, as Figure 6 shows. The smoking rate curve shows that the number of smokers decreased between 1996 and 2015, from 23% to 14% (a decrease of about 39.1 percentage points). Smoking rate has declined continuously in recent years. Figure 6 also shows that the incidence of lung cancer in men declined from 88 per 100,000 in 1996 to 58 per 100,000 in 2015, a reduction of 34.1 percentage points. As a result, lung cancer incidence had decreased along with smoking rate declining in men.

Figure 6. Relationship between smoking and lung cancer incidence, 1996-2015.



Accuracy of Deep Neural Network Models

Table 3 summarizes the performance of the 4 DNN models. The proposed models had good accuracy and area under the receiver operating characteristic curve (AUROC), using the whole group as a baseline to reveal lung cancer incidence in elderly patients. Accuracies were 96.2% (95% CI 0.530-0.751, *P*=.002) for age 65 years and older, 94.3% (95% CI 0.459-0.643, *P*=.015) for men 65 years and older, and 93.2% (95% CI 0.437-0.689, *P*=.003) for women 65 years and older, which were higher than the whole group accuracy of 92.7% (95% CI 0.223-0.525, *P*=.002). Moreover, AUROCs were 0.931 (95% CI 0.499-0.593) for age 65 years and older, 0.927 (95% CI 0.506-0.681) for men 65 years and older, and 0.926 (95% CI 0.543-0.782) for women 65 years and older, performing better than the whole group at 0.913 (95% CI 0.564-0.803). This proposal model efficiently output identified risk factors, which was timesaving.

Table 3. Performance of the 4 DNN models.

Model	Accuracy (95% CI)	AUROC[a] (95% CI)	*P* value[b]
≥65 years	0.962 (0.530-0.751)	0.931(0.499-0.593)	.002
Men ≥65 years	0.943 (0.459-0.643)	0.927 (0.506-0.681)	.015
Women ≥65 years	0.932 (0.437-0.689)	0.926 (0.543-0.782)	.003
All	0.927 (0.223-0.525)	0.913 (0.564-0.803)	.002

[a]AUROC: area under the receiver operating characteristic curve.

[b]*P*<.05 was considered to indicate statistical significance.

Discussion

Principal Findings

We developed, to our knowledge, the first deep learning classification model to quantitatively identify corresponding risk factors for lung cancer for stratified groups of elderly people. By leveraging the weights of the DNN models, we identified risk factors for lung cancer in the elderly and quantitatively analyzed the risk factors' degree of influence. The proposed DNN models of 4 groups (age ≥65 years, women ≥65 years, men ≥65 years, and the whole population) achieved good performance in identifying lung cancer risk factors, with accuracy ranging from 0.927 (95% CI 0.223-0.525, *P*=.002) to 0.962 (95% CI 0.530-0.751, *P*=.002) and AUROCs ranging from 0.913 (95% CI 0.564-0.803) to 0.931 (95% CI 0.499-0.593). The proposed models had a fast training speed and high accuracy and efficiency compared with logistic regression [18] and previous models for targeted identification of lung cancer risk factors [12,36-40].

In recent years, the deep learning method has been applied more frequently in lung cancer detection and prediction due to its advantages of high accuracy and fast computing speed. Hosny and colleagues used deep learning networks to predict mortality risk stratification of patients with NSCLC [22]. Cha and

colleagues found that a deep learning method had high diagnostic performance in detecting operable lung cancer with chest radiographs [20]. The DNN model, which we proposed to extract risk factors, could also be applied to provide intervention indicators for lung cancer prevention and carry out targeted intervention measures.

Through integrating multidisciplinary data, we the employed the DNN method to identify key lung cancer risk factors in the elderly. We computed quantitative weights of different risk factors in a stratified population to deduce their degrees of influence on lung cancer incidence. Our results showed that DNN models identified specific risk factors of targeted elderly people. People who were 65 years or older were more sensitive to time since quitting and smoking frequency, especially in men in this age group: smoking frequency was the leading causative risk factor for lung cancer in men 65 years and older. Time since quitting and smoking at least 100 cigarettes over a lifetime were the main risk factors for lung cancer in women 65 years and older. Men 65 years and older had the highest lung cancer incidence in these stratified groups. Lung cancer incidence decreased more obviously in men than in women with a decline in smoking rate. Cancer history played an important role in the incidence of lung cancer. Taking part in more physical activities to enhance physical quality might reduce lung cancer incidence [41,42]. Smoking-related factors (eg, smoking frequency, time since quitting, smoking at least 100 cigarettes) were important risk factors for lung cancer in elderly patients. Risk factors such as smoking-related factors, exercise, and cancer history were intervention indicators in preventing lung cancer. Tammemagi and colleagues found that smokers aged 65 to 80 years were a high-risk group who might benefit from low-dose CT lung cancer screening [43]. Chen and colleagues found that regional application of effective primary cancer prevention strategies on smoking, poor diet, and other modifiable risk factors had a vast potential to reduce the burden of cancer and disparities in China [9]. These suggested that interventional measures targeting the main risk factors might be possible to prevent lung cancer occurrence.

Comparison With Prior Work

Previously, researchers conducted several models to identify lung cancer risk factors [36-40]. Table 4 shows a comparison of our model with previous models. Compared with previous models, our proposed model identified risk factors for lung cancer in the elderly with high accuracy and AUROC. Our model used data from a larger population, more lung cancer occurrence-related risk factors, and a more efficient identification algorithm than previous models. Our DNN models had faster training speeds than previous models when training on the same scale of big data, which could save a lot of time. Moreover, we balanced and preprocessed the data before training the DNN models, which was helpful to improve model accuracy effectively.

Table 4. Comparison of our model with previous models for identifying lung cancer risk factors.

Model	Population	Method	Risk factors	Accuracy	AUROC[a]
Our model	235,673	Deep neural network	As listed in the Results section	0.927	0.913
Panayiotis, 2016 [36]	25,486	Dynamic Bayesian network	Demographics, smoking status, family history of cancer, cancer history, comorbidities related to lung cancer, occupational exposures, and low-dose computed tomography screening outcomes	0.65	0.75
Wang, 2019 [37]	961	Conditional Gaussian Bayesian network	Age, sex, level of education, region, urbanization, diagnosis-based factors, prior utilization factors, prescription factors	0.67	N/A[b]
Ankit, 2012 [38]	70,132	Decision tree	Age, birthplace, cancer grade, diagnostic confirmation, farthest extension of tumor, type of surgery performed, reason for no surgery, order of surgery and radiation therapy, scope of regional lymph node surgery	0.863	0.91
Xie, 2014 [39]	1703	Artificial neural network	41 risk factors: age, education level, marital status, income status, smoking, alcohol drinking, coffee intake, etc	0.838	N/A
Kaviarasi, 2019 [40]	321	Gaussian classifier	Age, sex, radiation sequence with surgery, first malignant primary indicator, radiation, etc	N/A	0.881

[a]AUROC: area under the receiver operating characteristic curve.

[b]Not available.

Some aspects of our results were similar to the results of these previous studies. In our results, smoking was the leading cause of lung cancer in the elderly. This view was consistent with the reported literature [2,44-46]. Nevertheless, we focused on some original findings in stratified groups of older people.

Limitations

This study had several limitations. First, we mainly focused on modifiable risk factors of lung cancer in the elderly. In the future, we should validate these identified modifiable risk factors using a simulated intervention process to prevent lung cancer. Second, because we used open survey data, we did not obtain the participants' genetic and dietary factors. We are matching the data to source region now and we will analyze lung cancer risk factors by region in the future.

Conclusions

This study demonstrated a quantitative method to identify risk factors for lung cancer in the elderly. The proposed models provided intervention indicators to prevent lung cancer,

especially in older men, which could be used with effective intervention methods to reduce lung cancer incidence in the elderly and improve their life quality in their later years. This approach might be used as a risk factor identification tool in other cancers and help physicians make decisions on cancer prevention.

Acknowledgments

This study was supported by the General Project on Humanities and Social Science Research of the Ministry of Education of China under grant no 19YJC870002, the National Key R&D Program of China under grant no 2016YFC0901602, and the Medical and Health Technology Innovation Project of the Chinese Academy of Medical Sciences under grant no 2019-I2M-2-002.

Conflicts of Interest

None declared.

References

1. U.S. National Library of Medicine. MedlinePlus. Non-small cell lung cancer. Bethesda, MD: U.S. Department of Health and Human Services, National Institutes of Health; 2019. URL: https://medlineplus.gov/ency/article/007194.htm [accessed 2019-06-20]
2. Schuller HM. The impact of smoking and the influence of other factors on lung cancer. Expert Rev Respir Med 2019 Aug;13(8):761-769. [doi: 10.1080/17476348.2019.1645010] [Medline: 31311354]
3. Park SK, Cho LY, Yang JJ, Park B, Chang SH, Lee K, Scientific Committee, Korean Academy of Tuberculosis and Respiratory Diseases. Lung cancer risk and cigarette smoking, lung tuberculosis according to histologic type and gender in a population based case-control study. Lung Cancer 2010 Apr;68(1):20-26. [doi: 10.1016/j.lungcan.2009.05.017] [Medline: 19545930]
4. Hahn EJ, Hooper M, Riker C, Butler KM, Rademacher K, Wiggins A, et al. Lung cancer worry and home screening for radon and secondhand smoke in renters. J Environ Health 2017;79(6):8-13 [FREE Full text] [Medline: 29135198]
5. Yang W, Zhao H, Wang X, Deng Q, Fan W, Wang L. An evidence-based assessment for the association between long-term exposure to outdoor air pollution and the risk of lung cancer. Eur J Cancer Prev 2016 May;25(3):163-172. [doi: 10.1097/CEJ.0000000000000158] [Medline: 25757194]
6. Lamm SH, Boroje IJ, Ferdosi H, Ahn J. Lung cancer risk and low (50 g/L) drinking water arsenic levels for US counties (2009-2013)-a negative association. Int J Environ Res Public Health 2018 Jun 07;15(6):1-21 [FREE Full text] [doi: 10.3390/ijerph15061200] [Medline: 29880761]
7. Cheng M, Chiu H, Tsai S, Chen C, Yang C. Calcium and magnesium in drinking-water and risk of death from lung cancer in women. Magnes Res 2012;25(3):112-119 [FREE Full text] [doi: 10.1684/mrh.2012.0318] [Medline: 23073359]
8. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Nikšić M, CONCORD Working Group. Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. Lancet 2018 Mar 17;391(10125):1023-1075 [FREE Full text] [doi: 10.1016/S0140-6736(17)33326-3] [Medline: 29395269]
9. Chen W, Xia C, Zheng R, Zhou M, Lin C, Zeng H, et al. Disparities by province, age, and sex in site-specific cancer burden attributable to 23 potentially modifiable risk factors in China: a comparative risk assessment. Lancet Glob Health 2019 Feb;7(2):e257-e269 [FREE Full text] [doi: 10.1016/S2214-109X(18)30488-1] [Medline: 30683243]
10. Lobach DF, Johns EB, Halpenny B, Saunders T, Brzozowski J, Del Fiol G, et al. Increasing complexity in rule-based clinical decision support: the symptom assessment and management intervention. JMIR Med Inform 2016 Nov 08;4(4):e36 [FREE Full text] [doi: 10.2196/medinform.5728] [Medline: 27826132]
11. 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 Aug 4;365(5):395-409 [FREE Full text] [doi: 10.1056/NEJMoa1102873] [Medline: 21714641]
12. Zahnd WE, Eberth JM. Lung cancer screening utilization: a behavioral risk factor surveillance system analysis. Am J Prev Med 2019 Aug;57(2):250-255. [doi: 10.1016/j.amepre.2019.03.015] [Medline: 31248742]
13. Moyer VA, U.S. Preventive Services Task Force. Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2014 Mar 4;160(5):330-338. [doi: 10.7326/M13-2771] [Medline: 24378917]
14. Berkowitz Z, Zhang X, Richards TB, Peipins L, Henley SJ, Holt J. Multilevel small-area estimation of multiple cigarette smoking status categories using the 2012 Behavioral Risk Factor Surveillance System. Cancer Epidemiol Biomarkers Prev 2016 Oct;25(10):1402-1410 [FREE Full text] [doi: 10.1158/1055-9965.EPI-16-0244] [Medline: 27697795]
15. Luna JM, Chao H, Diffenderfer ES, Valdes G, Chinniah C, Ma G, et al. Predicting radiation pneumonitis in locally advanced stage II-III non-small cell lung cancer using machine learning. Radiother Oncol 2019 Apr;133:106-112. [doi: 10.1016/j.radonc.2019.01.003] [Medline: 30935565]
16. Palani D, Venkatalakshmi K. An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification. J Med Syst 2018 Dec 18;43(2):21. [doi: 10.1007/s10916-018-1139-7] [Medline: 30564924]

17. Ahmed K, Emran AA, Jesmin T, Mukti RF, Rahman MZ, Ahmed F. Early detection of lung cancer risk using data mining. Asian Pac J Cancer Prev 2013;14(1):595-598 [FREE Full text] [doi: 10.7314/apjcp.2013.14.1.595] [Medline: 23534801]
18. Liu Y, Pleasants RA, Croft JB, Lugogo N, Ohar J, Heidari K, et al. Body mass index, respiratory conditions, asthma, and chronic obstructive pulmonary disease. Respir Med 2015 Jul;109(7):851-859 [FREE Full text] [doi: 10.1016/j.rmed.2015.05.006] [Medline: 26006753]
19. Lynch CM, Abdollahi B, Fuqua JD, de Carlo AR, Bartholomai JA, Balgemann RN, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int J Med Inform 2017 Dec;108:1-8 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.09.013] [Medline: 29132615]
20. Cha MJ, Chung MJ, Lee JH, Lee KS. Performance of deep learning model in detecting operable lung cancer with chest radiographs. J Thorac Imaging 2019 Mar;34(2):86-91. [doi: 10.1097/RTI.0000000000000388] [Medline: 30802232]
21. Schwyzer M, Ferraro DA, Muehlematter UJ, Curioni-Fontecedro A, Huellner MW, von Schulthess GK, et al. Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks - initial results. Lung Cancer 2018 Dec;126:170-173. [doi: 10.1016/j.lungcan.2018.11.001] [Medline: 30527183]
22. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. PLoS Med 2018 Nov;15(11):e1002711 [FREE Full text] [doi: 10.1371/journal.pmed.1002711] [Medline: 30500819]
23. US Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System.: U.S. Department of Health & Human Services; 2019. URL: https://www.cdc.gov/brfss/ [accessed 2020-01-27]
24. United States Environmental Protection Agency. Environmental Data database. 2019. URL: https://www.epa.gov/
25. Samet JM, Avila-Tang E, Boffetta P, Hannan LM, Olivo-Marston S, Thun MJ, et al. Lung cancer in never smokers: clinical epidemiology and environmental risk factors. Clin Cancer Res 2009 Sep 15;15(18):5626-5645 [FREE Full text] [doi: 10.1158/1078-0432.CCR-09-0376] [Medline: 19755391]
26. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc 2008 May;83(5):584-594 [FREE Full text] [doi: 10.4065/83.5.584] [Medline: 18452692]
27. Zheng S, Jabbour SK, O'Reilly SE, Lu JJ, Dong L, Ding L, et al. Automated information extraction on treatment and prognosis for non-small cell lung cancer radiotherapy patients: clinical study. JMIR Med Inform 2018 Feb 01;6(1):e8 [FREE Full text] [doi: 10.2196/medinform.8662] [Medline: 29391345]
28. Im Y, Park HY, Shin S, Shin SH, Lee H, Ahn JH, et al. Prevalence of and risk factors for pulmonary complications after curative resection in otherwise healthy elderly patients with early stage lung cancer. Respir Res 2019 Jul 04;20(1):136 [FREE Full text] [doi: 10.1186/s12931-019-1087-x] [Medline: 31272446]
29. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun 01;16(6):321-357. [doi: 10.1613/jair.953]
30. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999 Mar;8(1):3-15. [Medline: 10347857]
31. Kalman D. A singularly valuable decomposition: the SVD of a matrix. Coll Math J 1996 Jan;27(1):2. [doi: 10.2307/2687269]
32. The HDF group. HDF5 format. 2019. URL: https://support.hdfgroup.org/HDF5/
33. Caffe. Caffe deep learning framework. 2019. URL: http://caffe.berkeleyvision.org/
34. Jatoi I, Anderson WF, Miller AB, Brawley OW. The history of cancer screening. Curr Problems Surg 2019 Apr;56(4):138-163. [doi: 10.1067/j.cpsurg.2018.12.006]
35. Zhou H, Huang Y, Qiu Z, Zhao H, Fang W, Yang Y, et al. Impact of prior cancer history on the overall survival of patients newly diagnosed with cancer: a pan-cancer analysis of the SEER database. Int J Cancer 2018 Oct 01;143(7):1569-1577 [FREE Full text] [doi: 10.1002/ijc.31543] [Medline: 29667174]
36. Petousis P, Han SX, Aberle D, Bui AAT. Prediction of lung cancer incidence on the low-dose computed tomography arm of the National Lung Screening Trial: a dynamic Bayesian network. Artif Intell Med 2016 Sep;72:42-55 [FREE Full text] [doi: 10.1016/j.artmed.2016.07.001] [Medline: 27664507]
37. Wang K, Chen J, Wang K. Medical expenditure estimation by Bayesian network for lung cancer patients at different severity stages. Comput Biol Med 2019 Mar;106:97-105. [doi: 10.1016/j.compbiomed.2019.01.015] [Medline: 30708222]
38. Ankit A, Sanchit M, Ramanathan N, Lalith P, Aloi C. Lung cancer survival prediction using ensemble data mining on SEER data. Sci Programming 2012;20(1):29-42 [FREE Full text] [doi: 10.3233/SPR-2012-0335]
39. Xie N, Hu L, Li T. Lung cancer risk prediction method based on feature selection and artificial neural network. Asian Pac J Cancer Prev 2014;15(23):10539-10542 [FREE Full text] [doi: 10.7314/apjcp.2014.15.23.10539] [Medline: 25556505]
40. Kaviarasi R, Gandhi RR. Accuracy enhanced lung cancer prognosis for improving patient survivability using proposed Gaussian classifier system. J Med Syst 2019 May 24;43(7):201. [doi: 10.1007/s10916-019-1297-2] [Medline: 31127444]
41. Granger CL, McDonald CF, Berney S, Chao C, Denehy L. Exercise intervention to improve exercise capacity and health related quality of life for patients with Non-small cell lung cancer: a systematic review. Lung Cancer 2011 May;72(2):139-153. [doi: 10.1016/j.lungcan.2011.01.006] [Medline: 21316790]
42. Arbane G, Tropman D, Jackson D, Garrod R. Evaluation of an early exercise intervention after thoracotomy for non-small cell lung cancer (NSCLC), effects on quality of life, muscle strength and exercise tolerance: randomised controlled trial. Lung Cancer 2011 Feb;71(2):229-234. [doi: 10.1016/j.lungcan.2010.04.025] [Medline: 20541832]

43. Tammemägi MC, Church TR, Hocking WG, Silvestri GA, Kvale PA, Riley TL, et al. Evaluation of the lung cancer risks at which to screen ever- and never-smokers: screening rules applied to the PLCO and NLST cohorts. PLoS Med 2014 Dec;11(12):e1001764 [FREE Full text] [doi: 10.1371/journal.pmed.1001764] [Medline: 25460915]
44. Guo NL, Wan Y. Pathway-based identification of a smoking associated 6-gene signature predictive of lung cancer risk and survival. Artif Intell Med 2012 Jun;55(2):97-105 [FREE Full text] [doi: 10.1016/j.artmed.2012.01.001] [Medline: 22326768]
45. Giuliani ME, Liu G, Xu W, Dirlea M, Selby P, Papadakos J, et al. Implementation of a novel electronic patient-directed smoking cessation platform for cancer patients: interrupted time series analysis. J Med Internet Res 2019 Apr 09;21(4):e11735 [FREE Full text] [doi: 10.2196/11735] [Medline: 30964445]
46. Wraith D, Mengersen K. Assessing the combined effect of asbestos exposure and smoking on lung cancer: a Bayesian approach. Stat Med 2007 Feb 28;26(5):1150-1169. [doi: 10.1002/sim.2602] [Medline: 16779874]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BRFSS: Behavioral Risk Factor Surveillance System
Caffe: Convolutional Architecture for Fast Feature Embedding
CAT: computerized axial tomography
COPD: chronic obstructive pulmonary disease
CT: computed tomography
DNN: deep neural network
HDF5: hierarchical data format version 5
NSCLC: non–small cell lung cancer
SMOTE: synthetic minority oversampling technique

Edited by G Eysenbach; submitted 04.01.20; peer-reviewed by A Brown, W Clinton; comments to author 18.01.20; revised version received 19.01.20; accepted 22.01.20; published 17.03.20

Please cite as:
Chen S, Wu S
Identifying Lung Cancer Risk Factors in the Elderly Using Deep Neural Networks: Quantitative Analysis of Web-Based Survey Data
J Med Internet Res 2020;22(3):e17695
URL: http://www.jmir.org/2020/3/e17695/
doi: 10.2196/17695
PMID: 32181751