Original Article

An application of machine learning based on real-world data: Mining features of fibrinogen in clinical stages of lung cancer between sexes

Fangtao Yin[1#], Hongyu Zhu[1#], Songlin Hong[2], Chen Sun[2], Jie Wang[3,4], Mengting Sun[3,4], Lin Xu[1], Xiaoxiao Wang[5], Rong Yin[1,3,4]

[1]Department of Thoracic Surgery, The Affiliated Cancer Hospital of Nanjing Medical University & Jiangsu Cancer Hospital & Jiangsu Institute of Cancer Research, Jiangsu Key Laboratory of Molecular and Translational Cancer Research, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing, China; [2]F&E Data Technology (Tianjin) Corp., Tianjin, China; [3]Department of Scientific Research, The Affiliated Cancer Hospital of Nanjing Medical University & Jiangsu Cancer Hospital & Jiangsu Institute of Cancer Research, Jiangsu Key Laboratory of Molecular and Translational Cancer Research, Nanjing, China; [4]Biobank of Lung Cancer, Jiangsu Biobank of Clinical Resources, Nanjing, China; [5]GCP Research Center, Affiliated Hospital of Nanjing University of Chinese Medicine, Jiangsu Province Hospital of TCM, Nanjing, China

Contributions: (I) Conception and design: X Wang, R Yin; (II) Administrative support: None; (III) Provision of study materials or patients: F Yin, H Zhu, J Wang, M Sun, L Xu, X Wang, R Yin; (IV) Collection and assembly of data: F Yin, H Zhu, M Sun; (V) Data analysis and interpretation: F Yin, H Zhu, S Hong, C Sun; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Rong Yin. Department of Thoracic Surgery, The Affiliated Cancer Hospital of Nanjing Medical University & Jiangsu Cancer Hospital & Jiangsu Institute of Cancer Research, Jiangsu Key Laboratory of Molecular and Translational Cancer Research, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing, China. Email: rong_yin@njmu.edu.cn; Xiaoxiao Wang. GCP Research Center, Affiliated Hospital of Nanjing University of Chinese Medicine, Jiangsu Province Hospital of TCM, Nanjing, China. Email: wxx1201@hotmail.com.

Background: Lung cancer is the most threatening malignant tumor to human health and life. Using a variety of machine learning algorithms and statistical analyses, this paper explores, discovers and demonstrates new indicators for the early diagnosis of lung cancer and their diagnostic performance from large samples of clinical data in the real world.

Methods: By applying machine learning methods, including minimum description length (MDL), naive Bayesian (NB), K-means (KM), nonnegative matrix factorization (NMF), and decision tree (DT), based on large sample data of 2,502 patients, we built a classification model and systematically explored differences in fibrinogen levels in different clinical stages of lung cancer between the sexes. We also validated the reliability of the model by testing it on a validation cohort of 447 patients. This report adheres to the "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis" (TRIPOD) statement for the reporting of prediction models.

Results: The analysis revealed significant differences in fibrinogen levels, pleural effusion, chlorine levels, A-G ratio, glutamic-oxaloacetic transaminase and alkaline phosphatase levels as well as in sex composition between the early-stage lung cancer group and the middle-late-stage lung cancer group. The classification model created by the combination of fibrinogen, alkaline phosphatase and sex demonstrated good performance with an AUC of 73.5%. In addition, in males, a fibrinogen level of 2.94 g/L could initially serve as the upper limit for determining the early-stage lung cancer group, but a level of 3.91 g/L could be preliminarily used as a reference threshold for the lower limit for middle- to late-stage lung cancer. This latter level could also serve as the upper limit of the critical value for early-stage lung cancer in females.

Conclusions: An integrated application based on supervised and unsupervised machine learning algorithms could effectively explore the potential links contained in the clinical data and reveal the differences in fibrinogen levels in different clinical stages of lung cancer between the sexes, which could provide a new reference basis for lung cancer staging.

Keywords: Machine learning; lung cancer; early diagnosis; fibrinogen; sex

Submitted Jun 15, 2020. Accepted for publication Feb 04, 2021.
doi: 10.21037/atm-20-4704
View this article at: http://dx.doi.org/10.21037/atm-20-4704

Introduction

Lung cancer is one of the most threatening malignant tumors to human health and life; its mortality rate is the highest among malignant tumors in the United States (1,2). The prevention and treatment of lung cancer is a difficult problem faced worldwide. Precise treatment for lung cancer patients is also a hot topic of current research, particularly for different sexes. Significant sex differences in the morbidity rate, pathological type, pathogenesis, treatment and prognosis of lung cancer have been noted (3). Therefore, deeply exploring the clinical characteristics and medical mechanisms of lung cancer between the sexes would provide a theoretical basis for the precise treatment of this disease.

Based on real-world data on malignant lung tumors and the application of machine learning algorithms, this paper explored the clinical variation in the levels and characteristics of fibrinogen combined with other relevant indicators in two clinical stages of lung cancer between the sexes.

First, the minimum description length (MDL) algorithm was used to select and explore the characteristics of the data. Then, based on medical knowledge, fibrinogen levels and sex were screened as clinical indicators to distinguish the differences between lung cancer stages. Further statistical tests were used to verify the reliability of the results. The naive Bayes (NB) classification model was created from the combination of fibrinogen level, alkaline phosphatase and sex. The reliability of this model was validated using two methods in a validation cohort. Finally, nonnegative matrix factorization (NMF) and decision tree (DT) algorithms were used to explore the main characteristic expressions of fibrinogen with sex in the two stages of lung cancer.

These combined methods completely demonstrated the process of real-world clinical data research with machine learning algorithms. In addition, the different characteristics of fibrinogen based on sex between the two stages of lung cancer were also revealed, potentially providing a new reference for the diagnosis of lung cancer staging. The workflow of this study is shown in *Figure 1A*. We present the following article in accordance with the TRIPOD reporting checklist (available at http://dx.doi.org/10.21037/atm-20-4704) (4).

Methods

Source of data and participants

Based on the application of machine learning and data mining technology, we aimed to explore the relationship between patients' clinical characteristics and pathological stage of cell lung cancer in real-world data that could then be used to develop a prediction model.

We only used data from our thoracic surgery medical center in this study. From January 1, 2013 to July 17, 2019, we retrospectively collected data from electronic medical records (EMRs) on every lung cancer patient admitted to our medical wards. A total of 3588 patients admitted from January 1, 2013 to May 31, 2018 were recruited in the training cohort, whereas 1,417 patients admitted from June 1, 2018 to July 17, 2019 were recruited in the validation cohort. In the training cohort, 1,086 patients were excluded due to the presence of benign tumors, readmission, or lack of pathological or clinical information. In addition, 970 were excluded for the same reasons in the validation cohort.

The study was conducted according in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the ethics committee of Jiangsu Cancer Hospital (IRB approval: 2017-K-001). As a retrospective study without any effects on individual patient outcomes, the ethical review board waived the need for individual patient informed consent.

Outcomes

The primary outcome was whether the pathological stage of lung cancer patients was stage I or stage II, III, or IV. Three thoracic surgeons restaged the disease based on the patient's clinical and pathological information recorded in our EMR

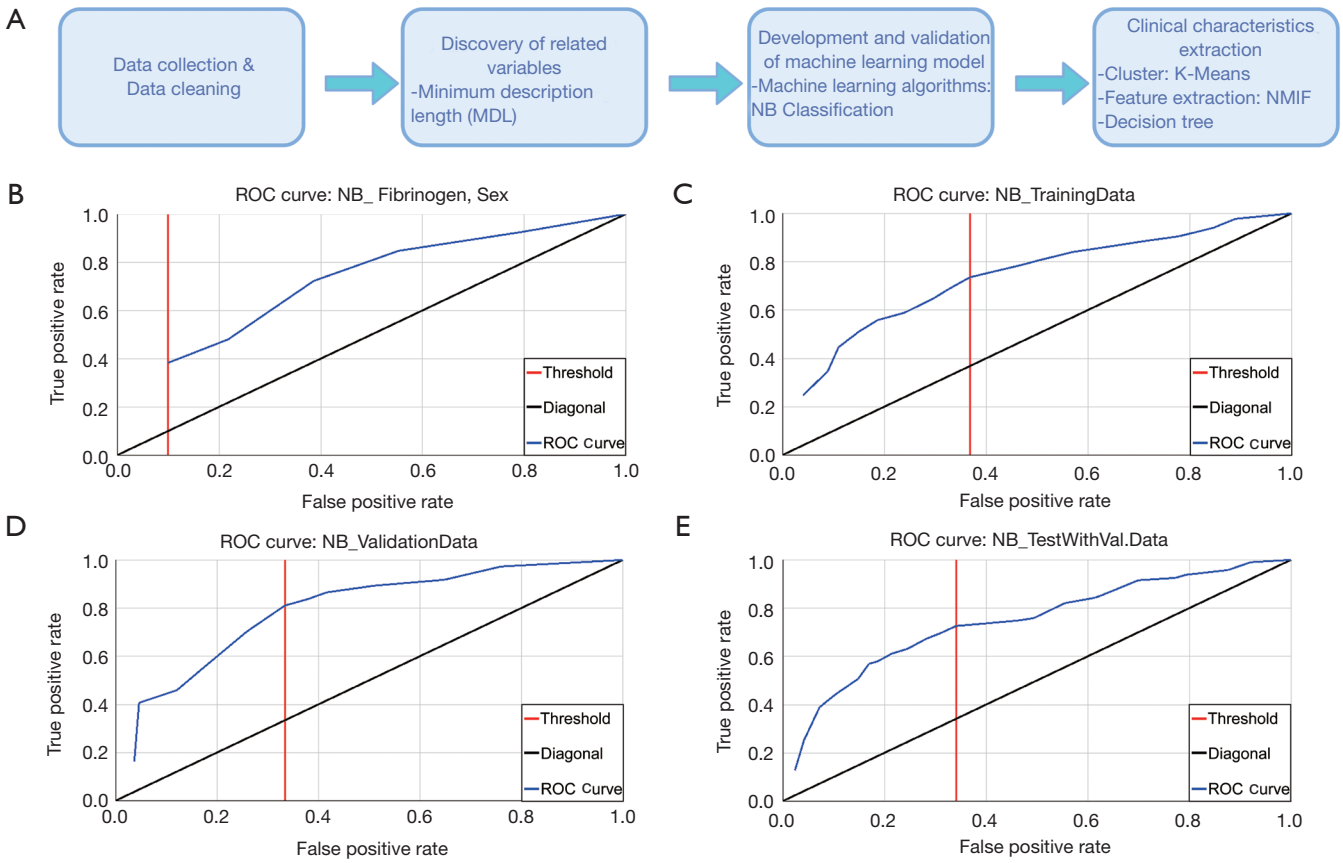

Figure 1 Overview of study workflow and ROC curves of the machine learning model. (A) Overview of the study workflow. First, the clinical data were collected from Jiangsu Cancer Hospital, and necessary data cleaning was performed. The minimum description length (MDL) algorithm was used to select and explore data characteristics. Then, based on medical knowledge, fibrinogen, alkaline phosphatase and sex were screened as clinical indicators to distinguish the differences between the stages of lung cancer. Further statistical tests were used to verify the reliability of the results. Finally, nonnegative matrix factorization (NMF) and decision tree (DT) algorithms were used to explore the main characteristic expressions of fibrinogen with sex in the two stages of lung cancer. (B) NB model with variables fibrinogen and sex (NB_ Fibrinogen, Sex). AUC, 0.715. (C) NB model with variables fibrinogen, sex and alkaline phosphatase (NB_TrainingData). AUC, 0.735. (D) NB model with the same variables used in the training model, sex, fibrinogen, and alkaline phosphatase, built from the validation data (NB_ValidationData). AUC, 0.79. (E) NB model established by the training data tested with the validation data (NB_TestWithVal.Data). AUC, 0.74. AUC, area under the curve.

database according to the TNM 8th edition.

Predictors

Our EMR database consists of 227 elements, such as patients' demographics and clinical features (e.g., diagnoses, medications, and laboratory results), from admission to discharge. These characteristics were analyzed and screened as potential predictors. We also performed a series of statistical tests, including correlation tests, between potential predictors and the target variable and analyses of potential predictor differences between the two target groups.

Sample size and missing data

All available admission data from our cohort were used without a priori sample size calculations.

At the beginning of the analysis, we first performed necessary data cleaning processes, such as identifying

missing values and data inconsistency processes. According to several variables with high correlation, null value filtering was performed to ensure validity, and null value data in the data set were deleted. Although a certain amount of data was reduced, the rigor and reliability of the analysis results were guaranteed to the greatest extent.

Statistical analysis methods

In the training cohort, 227 demographic and clinical indicators were comprehensively analyzed to determine differences between the two groups, and the MDL algorithm was used to select and explore the characteristics of the data. Seven nontumor marker variables were selected as candidate predictors to establish the classification model. A series of statistical tests were then performed to verify the results obtained from machine learning algorithms.

After the discovery of related variables, we applied the classic machine learning algorithm naive Bayes (NB) classification based on probability theory to establish classification models to evaluate related variables for distinguishing early- and middle- to late-stage lung cancer. By treating fibrinogen level and sex as the basic variables in the classification models, we combined pleural effusion, chlorine level, A-G ratio, glutamic-oxaloacetic transaminase and alkaline phosphatase levels to create the different classification models. Then, the top two best performing models were chosen.

These models' positive (middle- to late-stage lung cancer) and negative (early-stage lung cancer) predictive rates, area under the curve (AUC) values, sensitivities and specificities were calculated to assess the performance of the models.

The validation data were used as an independent dataset to establish a Bayesian classification model with the abovementioned features (NB_ValidationData). Then, we compared the modeling results with those obtained from the training data (NB_TrainingData) to evaluate the performance of the model. In addition, the validation data were also used to verify the previous model established with the training data (NB_TestWithVal. Data) to assess the reliability of the prediction model from another aspect.

The K-Means algorithm, nonnegative matrix factorization (NMF) and decision tree (DT) algorithms were then used to explore the common features between the characteristics selected and the stage of lung cancer.

Results

Participants

Training data were collected from Jiangsu Cancer Hospital over a 5-year period from January 2013 to May 2018. These data contained information, including patients' demographic information, admission, laboratory tests, medical examination, surgical information, medical imaging, pathological information and other multidimensional data. A total of 227 variables were collected from 2,502 patients.

These malignant lung tumor patients mainly included 1,393 adenocarcinoma patients and 329 patients with squamous cell carcinoma (SCC). We divided these 2,502 patients into two groups according to the International TNM Staging of Lung Cancer: Group A had a sample size of 1,561 stage I lung cancer patients (940 adenocarcinoma and 122 SCC). Group B had a sample size of 941 stage II, III and IV lung cancer patients (453 adenocarcinoma and 207 SCC). At the beginning of the analysis, we first performed necessary data cleaning processes, such as identification of missing values and data inconsistency processes.

The validation cohort was collected from Jiangsu Cancer Hospital from May 2018 to July 2019 and included a sample size of 447. The data were processed in the same way as the training cohort. The data summary is shown in *Table 1*.

The top 7 nontumor marker variables were selected as candidate predictors to establish classification models and verified by statistical testing

First, we took label A/B as the target variable to explore the difference between these two groups, assigning a value of 0 to Group A and a value of 1 to Group B. Then, the 227 demographic and clinical indicators were comprehensively analyzed to determine the differences between the two groups. By applying the MDL algorithm to variable selection, we treated each attribute in a given dataset as a simple predictive model for the target. Then, all models were compared and graded using their corresponding MDL scores.

On the premise of combining medical knowledge, we deleted some obvious related variables, such as distant metastasis location on CT and remote metastasis on MRI. As recommended by MDL, we selected the indicators with importance values greater than 0 (as shown in *Table 2*) as the predictive candidate indicators to explore the distinctions

Table 1 Characteristics of the study population: training cohort (n=2,502) and validation cohort (n=447)

Variables	Categories	Training cohort (n=2502)		Validation cohort (n=447)	
		Stage I (n=1,561)	Stage II, III and IV (n=941)	Stage I (n=336)	Stage II, III and IV (n=111)
Sex	Male	767	676	147	73
	Female	794	265	189	38
Age	Median	61	63	60	64
	Range	28–86	25–87	30–80	43–82
Tobacco use	Yes	233	365	19	15
	No	376	430	313	93
	Quit	0	0	2	3
	Unknown	952	146	2	0
Histological type	Adenocarcinoma	940	453	216	58
	Squamous cell carcinoma	122	207	10	28
	Others	499	281	110	25
Fibrinogen	Median	2.71	3.22	2.79	3.39
	Range	1.19–5.76	0.93–5.96	1.3–5.4	1.23–5.58
Pleural effusion	Yes	194	223	2	4
	No	281	411	63	15
	Null	1,086	307	271	92
Chlorine	Median	103	102	103.3	102.85
	Range	82–118	82–112	87.7–116.1	91.6–112.3
Albumin-globulin (A/G) ratio	Median	1.7	1.585	1.7	1.5
	Range	1.04–2.9	0.64–2.8	1–2.6	0.8–2.5
Glutamic-oxaloacetic transaminase	Median	21	22	20	22
	Range	6–244	8–257	0.8–162	1–746
Alkaline phosphatase	Median	71	79	65	75
	Range	14–319	4–1,449	27–291	43–215
Albumin	Median	43	42	34.15	34.65
	Range	25–54	22–55	2.33–52.1	1.32–54.1
Monocytes	Median	0.445	0.46	7.47	7.605
	Range	0–1.74	0–4.46	3.01–30.32	3.75–21.79
Initial symptoms	Hemoptysis	5	18	1	2
	Coughing	72	154	45	39
	Physical findings	281	180	272	55
	Expectoration	3	5	0	1
	Chest pain	9	14	6	5
	Others symptoms/ null	1,191	570	12	9
High-density lipoprotein	Median	1.32	1.27	1.11	1.08
	Range	0.72–2.87	0.53–3.66	0.51–2.62	0.57–1.94

between early- and middle- to late-stage lung cancer. By analyzing *Table 2*, we found that the importance values of fibrinogen level and sex were greater than those of the remaining important variables.

The top 7 nontumor marker variables were selected to establish a classification model, but they had different numbers of valid cases. For example, fibrinogen level had 1,816 valid cases, and sex had 2,502 valid cases.

To verify the results gained from the machine learning algorithms, we performed a series of statistical tests, including correlation tests between fibrinogen level/sex and the target variable and analyses of fibrinogen level/sex differences between the two target groups. The statistical results were consistent with the machine learning results.

The Spearman correlation coefficient between fibrinogen level and the target variable was 0.303 ($P<2.2\times10^{-16}$), and the correlation coefficient between sex and the target variable was –0.233 ($P<2.2\times10^{-16}$). These results indicated a significant relationship between fibrinogen level/sex and lung cancer stage. The comparisons of fibrinogen level/sex difference between the two target groups indicated a significant difference with a P value approximately equal to 0. Other major related variables, including pleural effusion, chlorine level, A-G ratio, glutamic oxaloacetic transaminase and alkaline phosphatase levels, were also subjected to the above statistical tests, and the results were similar, indicating a significant correlation in distinguishing early- and middle- to late-stage lung cancer. This finding demonstrated the reliability of the results that were obtained by the MDL algorithm.

Table 2 Table of candidate indicators

Variable	Order	Importance
Fibrinogen*	1	0.039434316
Sex*	2	0.03285276
Pleural effusion*	3	0.005590354
Chlorine*	4	0.005446061
Albumin-globulin (A/G) ratio*	5	0.003882843
Glutamic-oxaloacetic transaminase*	6	0.00385326
Alkaline phosphatase*	7	0.003628277
Lymphatic metastasis (CT)	8	0.002124314
Tumor (X-ray)	9	0.001673887
Albumin	10	0.001413229
Monocytes	11	0.001240597
Initial symptoms	12	0.001149529
Blood coagulation	13	0.001130126
High-density lipoprotein	14	0.001040269

As recommended by the MDL algorithm, indicators with importance values greater than 0 were selected as predictive candidate indicators to explore the distinctions between early- and middle- to late-stage lung cancer. *, the top 7 nontumor marker variables are shown.

The performance of the machine learning model reached a good level

After the discovery of related variables, we applied NB classification based on probability theory to establish classification models to evaluate related variables for distinguishing early- and middle- to late-stage lung cancer. By treating fibrinogen level and sex as basic variables, we combined pleural effusion, chlorine level, A-G ratio, glutamic-oxaloacetic transaminase and alkaline phosphatase levels to build different classification models. Then, the top two best performing models were chosen. The positive (middle- to late-stage lung cancer) and negative (early-stage lung cancer) predictive rates, area under the curve (AUC) values, sensitivities and specificities are shown in *Table 3*.

Among the above models, the NB classification model generated from the basic composition of fibrinogen level and sex showed good performance. The positive predictive rate of the NB model was relatively high (72–73%), and the

Table 3 Summary statistics for lung cancer stage classification models

Variables	Algorithm	Average accuracy	Positive predictive rate	Negative predictive rate	AUC	Sensitivity	Specificity
Sex, fibrinogen	Naive Bayesian	67.4%	72.4%	62.5%	71.5%	72.4%	37.5%
Sex, fibrinogen, alkaline phosphatase	Naive Bayesian	67.8%	65.3%	70.3%	73.5%	73.5%	37.1%

AUC, area under the curve.

Table 4 Model validation

Variable	Average accuracy	Positive predictive value	Negative predictive value	AUC	Sensitivity	Specificity
NB_TrainingData	67.77%	65.29%	70.25%	0.73	73.53%	62.88%
NB_ValidationData	66.38%	54.05%	78.70%	0.79	81.08%	66.67%
NB_TestWithVal.Data	69.55%	67.37%	71.72%	0.74	72.63%	65.86%

NB_TrainingData, NB model with variables fibrinogen, sex and alkaline phosphatase; NB_ValidationData, NB model with the same variables used in the training model, sex, fibrinogen, and alkaline phosphatase, built from the validation data; NB_TestWithVal.Data, NB model established by the training data tested with the validation data; AUC, area under the curve.

difference between the positive predictive rate and negative predictive rate within the model was approximately 10% (*Figure 1B*). On the other hand, the classification model that included fibrinogen level, sex and alkaline phosphatase performed better in predicting the target variable with a positive predictive rate of 65.3% and a negative predictive rate of 70.3%. This model exhibited the highest AUC of 73.5% among all other models and a relatively high sensitivity (73.5%) (*Figure 1C*). Given that all the models had lower specificities (25–37%), the differences were not significant.

The results showed that the best NB model, which included the variables of fibrinogen level, sex and alkaline phosphatase levels, had a prediction confidence of 35.5%, indicating that the performance of the model was 35.5% higher than that of a random predictive model. The AUC value of the model was 0.735, indicating a good level of performance.

The results suggested that the classification model built by the variables of sex, fibrinogen levels and alkaline phosphatase levels exhibited good predictive accuracy and potential clinical application value while verifying the different expression characteristics of sex, fibrinogen level and alkaline phosphatase levels in early and middle-late lung cancer. This information could provide a new reference basis for staging in the diagnosis of lung cancer in addition to tumor markers.

The machine learning model exhibits good performance with the validation cohort

By collecting validation data from May 2018 to July 2019 from 447 patients, we validated the efficiency of the Bayesian classification model.

First, the validation data were used as an independent dataset to establish a Bayesian classification model with the same features described above (NB_ValidationData). By comparing the modeling results with those from the model built from the training data (NB_TrainingData), the new model (*Figure 1D*) exhibited results similar to those of the previous model (*Figure 1C*), indicating that the related variables that were previously found by the machine learning algorithms can achieve good prediction results for different sources of data (*Table 4*).

In addition, the validation data were used to verify the previous model established by training data (NB_TestWithVal. Data), demonstrating that the results were satisfactory (*Figure 1E*), the model performed well and the reliability of the model was verified (*Table 4*).

Based on these two aspects of verification, we concluded that the model we established was efficient in predicting the target variables with satisfactory accuracy, indicating the reliability of the modeling algorithms.

The K-means (KM) algorithm divided fibrinogen into 7 categories

We explored and verified the differential nontumor marker variables in distinguishing early-stage and middle-late–stage lung cancer. Next, we further explored the characteristic expression of these nontumor marker variables in early-stage and middle- to late-stage lung cancer.

We applied the K-Means algorithm, which has been widely verified in clustering the key variables, to explore fibrinogen clustering patterns. The clustering results showed that the KM algorithm divided fibrinogen into 7 categories (*Table 5*).

These results accurately reflected the objective distribution of fibrinogen in this lung cancer population. The confidence of clustering ranged from 80.3% to 100%, and the average confidence was 91.94%. Clustering levels 1, 2 and 3 had a relatively low level of fibrinogen (1.255–3.275),

Table 5 The clustering results showed that the KM algorithm divided fibrinogen into 7 categories

ID	Lower limit	Upper limit	Confidence (%) =100.00	Support (n)
1	1.255	2.265	95.97	262
2	2.265	2.77	85.63	280
3	2.265	3.275	100	657
4	3.275	3.78	100	214
5	3.78	4.285	80.28	114
6	4.285	4.79	81.72	76
7	4.79	5.8	100	110

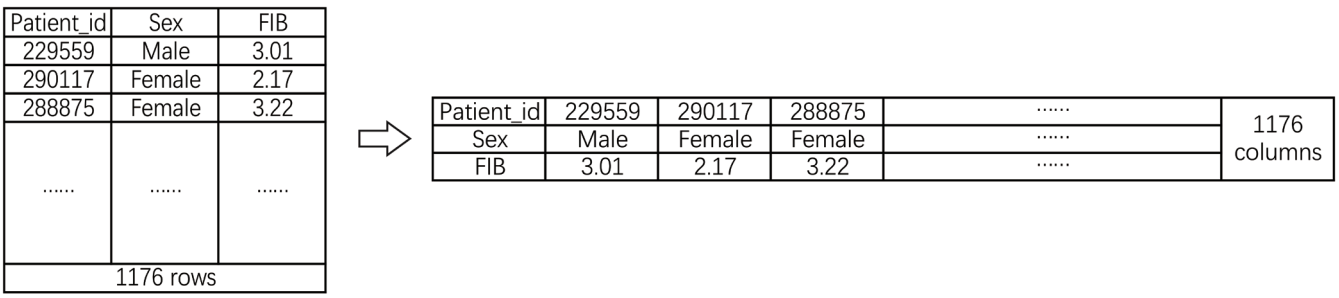

Figure 2 Schematic diagram of the NMF used to reduce the dimensionality of the data. NMF was used as a data-reduction technique to reduce the dimensions of the data in the early-stage lung cancer group. For this group, three variables were included, namely, patient ID, sex (male/female) and discretized fibrinogen (from category 1 to 7), with a total sample size of 1,176 that formed a nonsparse matrix. NMF, nonnegative matrix factorization.

4 and 5 were medium (3.275–4.285), and 6 and 7 were high (4.285–5.8).

Using these accurate clustering results, the fibrinogen level was carefully discretized, and combined with sex. We could explore the expression characteristics of different stages of lung cancer.

Nonnegative matrix factorization (NMF) for feature extraction

NMF is a feature extraction algorithm that is often used to analyze high-dimensional data or data with predictive variables that are not highly correlated to the target variables. By combining attributes, the NMF algorithm can produce meaningful patterns or topics. It compresses multivariate data by creating a user-defined number of features, each of which is a linear combination of the original set of attributes with nonnegative coefficients. The principle of the NMF algorithm can be simply described as follows: for any given nonnegative matrix V, NMF can find a nonnegative matrix W and a nonnegative matrix H to satisfy V=W*H, thereby decomposing a nonnegative matrix into the product of two nonnegative matrices.

To analyze the data in a more reliable manner, we deleted data with missing values directly instead of processing them with imputed missing data. Therefore, there were 1,176 samples in the early-stage lung cancer group and 639 samples in the middle- to late-stage lung cancer group for feature expression processing. NMF was used as a data-reduction technique to reduce the dimension of the data in the early-stage lung cancer group. For this group, there were three variables, including patient ID, sex (male/female) and discretized fibrinogen (from category 1 to 7), with a total sample size of 1,176, which formed a nonsparse matrix (*Figure 2*).

NMF first transposed the original 1,176 by 3 matrix into a new matrix of 3 by 1,176 (1,176 columns and 3 rows). In this new matrix, the first row represents patient ID, the second row represents sex, the third row represents fibrinogen category, and each column represents an

Table 6 NMF algorithm results

Stage	Feature 1		Feature 2	
	Feature	Coefficient	Feature	Coefficient
Early-stage lung cancer	Sex = M	0.867523	Sex = F	1.009724
	Fib. level =2	0.243168	Fib. level =2	0.385596
	Fib. level =3	0.203545	Fib. level =3	0.251904
	Fib. level =1	0.190356	Fib. level =1	0.166057
	Fib. level =4	0.101417	Fib. level =4	0.079301
Middle- to late-stage lung cancer	Sex = M	1.177668	Sex = F	0.5871693
	Fib. level =7	0.247422	Fib. level =2	0.2327963
	Fib. level =3	0.198573	Fib. level =3	0.162359
	Fib. level =4	0.153598	Fib. level =1	0.1020188
	Fib. level =2	0.120517	Fib. level =4	0.0931605
	Fib. level =6	0.102849		
	Fib. level =5	0.091088		

Two expression features were extracted from the early-stage and middle- to late-stage lung cancer groups. M, male; F, female; Fib. Level,

individual patient. We applied the NMF algorithm to reduce these 1,176-dimensional data based on patient sex and fibrinogen category to obtain the main sex and fibrinogen characteristics among early-stage lung cancer patients.

Then, we applied the same method to the middle- to late-stage lung cancer group with a sample size of 639, and the results are shown in *Table 6*.

The table shows the two expression features that were extracted from the early-stage lung cancer group. Feature 1 indicated that the fibrinogen level of males was relatively evenly distributed among lower fibrinogen categories 1, 2 and 3 (characteristic coefficients 0.19, 0.24 and 0.20, respectively) with a small part distributed in middle fibrinogen category 4 (characteristic coefficient 0.10). Feature 2 indicated that the fibrinogen level of females was mainly distributed among lower fibrinogen categories 1, 2 and 3 (characteristic coefficients 0.17, 0.39, and 0.25, respectively) as well. However, the distribution of these three lower levels was less balanced than that in Feature 1, and a small part was distributed in middle category 4 (characteristic coefficient 0.08). In general, the fibrinogen distribution was relatively similar between males and females in the early-stage lung cancer group with the majority of the data distributed among low levels of fibrinogen. The distribution of fibrinogen among categories 1 to 3 was more balanced in males, whereas female patients had a greater distribution of fibrinogen in category 2 compared with the other categories.

Additionally, two features were extracted from the middle- to late-stage lung cancer group. Feature 1 showed a widely distributed fibrinogen level among categories 2, 3, 4, 5, 6, and 7 in males (characteristic coefficients of 0.12, 0.20, 0.15, 0.09, 0.10, 0.25). Feature 2 indicated that the fibrinogen levels of females were still mainly distributed among the middle and lower categories 1, 2, 3 and 4 (characteristic coefficients 0.10, 0.23, 0.16 and 0.09), which is similar to the early-stage lung cancer group. The distribution of fibrinogen was substantially different between males and females in the middle-late lung cancer group. The distribution of fibrinogen in males was wide and mainly concentrated in category 7, whereas the distribution in females was narrower and more concentrated in lower fibrinogen categories, especially category 2.

Fibrinogen value as a predictor of lung cancer clinical stages between sexes by using the decision tree (DT) classification model

Next, we applied another machine learning algorithm, DT, to further explore related variables. The results showed that the DT classification model established by the sex and fibrinogen variables had an average predictive accuracy of 65.1%, an AUC value of 71.9%, a sensitivity of 78.7%, and

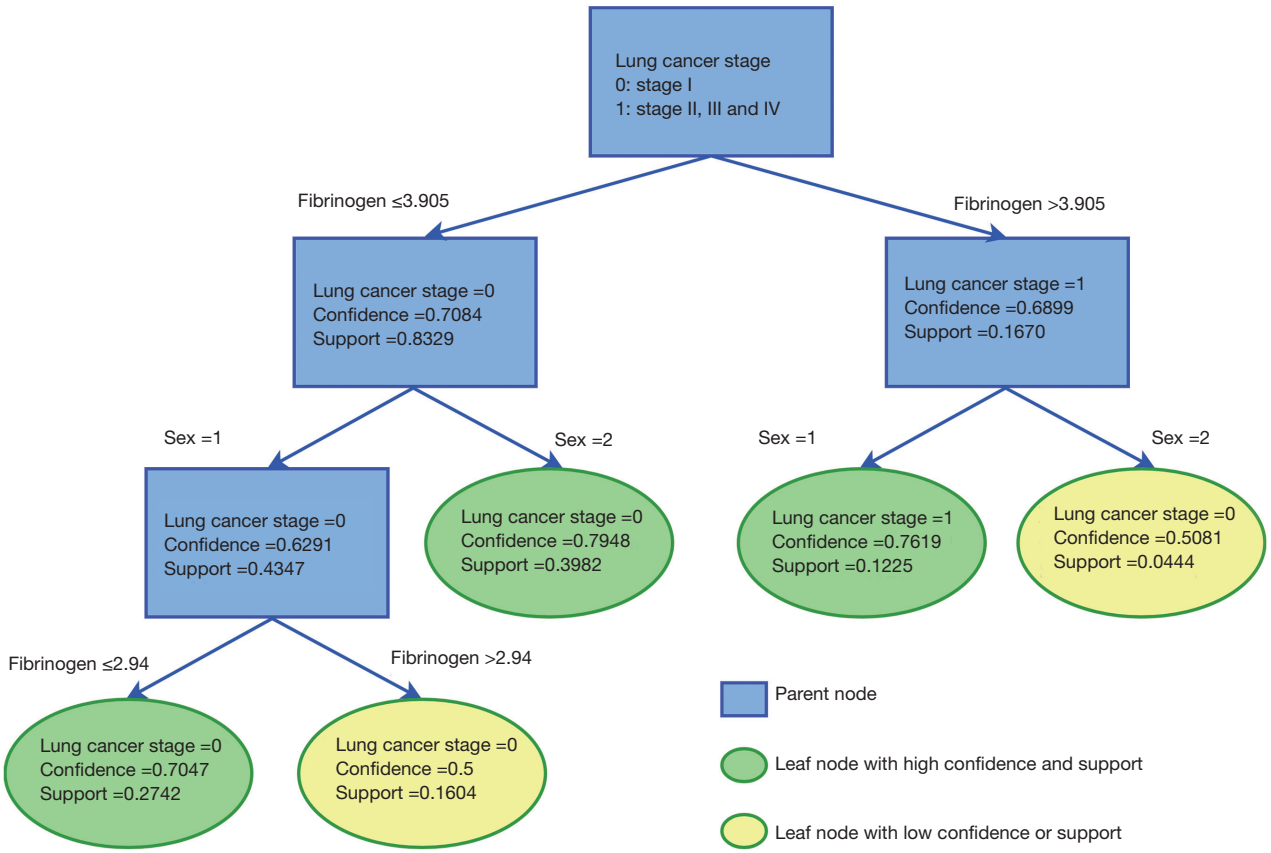

Figure 3 Decision tree (DT) classification model for lung cancer stage. DT classification model established by variables sex and fibrinogen. Lung cancer stage 0: stage I; 1: stage II, III and IV. Sex 1: male; 2: female. Rectangle box: parent node. Oval box: leaf node; green: leaf node with high confidence and support; yellow: leaf node with low confidence or support.

a specificity of 48.5%. The DT model rules are shown in *Figure 3*.

The model determined a threshold value of fibrinogen of 3.91. When the fibrinogen level was less than 3.91. The confidence and support values in predicting patients as having early-stage lung cancer were 70.8% and 83.3%, respectively. When fibrinogen was greater than 3.91, the confidence and support values in predicting patients as having middle- to late-stage lung cancer were 69.0% and 16.7%, respectively.

As shown in *Figure 3*, 2.94 could serve as the upper limit for predicting early-stage lung cancer in males. In addition, 3.91 could serve as the upper limit for early-stage lung cancer in females and as the lower limit of middle- to late-stage lung cancer in males.

In the above decision tree model, the distribution differences of fibrinogen between the early-stage lung cancer group and the middle- to late-stage lung cancer group for the sexes were clearly demonstrated, which was consistent with the results of the NMF algorithm in exploring the distribution characteristics of fibrinogen between the two groups of stages of lung cancer for the sexes.

Discussion

In lung cancer, clinically relevant prognostic information is provided by staging, which forms the basis for the treatment options. Although greater than 100 lung cancer prognostic or predictive markers have been published, including serum markers (CEA, CA-125, CYFR A 21-21, chromogranin A, NSE, and RBP), molecular markers (KRAS, EGFR, MET, and P53), most do not reach clinical implementation (5). Moreover, predictive or diagnostic effects using a single marker are usually limited and weak for lung cancer staging. Our study suggested that the classification model built from

sex, fibrinogen and alkaline phosphatase, which could easily be collected from EMRs, exhibited good predictive accuracy and potential clinical application value in distinguishing early lung cancer and middle-late lung cancer. Specifically, we provide a more precisive decision tree considering both fibrinogen level and sex for lung cancer staging that does not predict with a single marker and a single threshold. Obviously, it will be more effective to use artificial intelligence technologies to comprehensively establish a predictive model or system rather than a single biomarker. This could provide a new reference for staging the diagnosis of cell lung cancer and assist clinicians in diagnosis and decision-making before surgery and other subsequent therapies.

From the previous analysis, we can observe that as the lung cancer stage increases, fibrinogen also increases, especially in the male group and typically in a linear fashion. First, we will discuss the relationship between fibrinogen and malignant tumors.

The relationship between fibrinogen level and malignant tumors

Fibrinogen is an essential constituent of the coagulation system. Fibrinogen is mainly synthesized in the liver and released into the circulation in response to systemic inflammation and malignancy (6). Currently, it is believed that increasing fibrinogen levels in malignant tumor patients are mainly the result of the interaction between the tumor and the coagulation system. Tumor cells directly interact with endothelial cells and platelets and release bioactive substances, which promote the adhesion and aggregation of platelets and fibrinogen. In addition, tumor cells can secrete cancer procoagulant (CP) and tissue factor (TF) and a variety of fibrinolytic and antifibrinolytic substances, leading to activation of the blood coagulation system and contributing to thrombus formation. Tumor cells also induce endothelial cells to secrete fibrinolysis enzymes that activate inhibitors to stop the degradation of fibrinogen. When this effect is greater than the degradation of fibrinogen, the level of fibrinogen increases (7). The increase in plasma fibrinogen levels may lead to disordered coagulation function, which is closely related to the occurrence, development, recurrence and metastasis of malignant tumors (8). An increasing body of evidence has indicated the association between fibrinogen and tumor clinical stage, angiogenesis, metastatic spread, and response to therapy in patients with cancer (9). Therefore, fibrinogen can be used as an important indicator to monitor the progression of malignant tumor metastasis, therapeutic efficacy and prognosis.

Currently, few anticoagulants are used directly in the treatment of malignant tumors, but there is a consensus that anticoagulants can be used to treat cancer-related thrombosis. ASCO guidelines strongly recommend the use of low-molecular-weight heparin (LMWH) for the treatment of cancer - associated thrombosis (CAT) (10). A number of clinical trials of oral anticoagulants in the treatment of cancer-related thrombosis are on-going. On the other hand, many studies have shown that platelets play an important role in promoting tumor development and metastasis (11-13), and some have shown that inhibition of platelet targets, such as GPVI, can cause intratumor hemorrhage (14), thus achieving antitumor effects. We believe targeting the coagulant pathway might represent a promising strategy for cancer therapy.

Characteristics of fibrinogen in different lung cancer stages

By analyzing the results of clustering, feature expression analysis and decision tree classification performed in this paper, we can conclude that the level of fibrinogen in the early lung cancer male group would be evenly distributed among relatively low values (1.26–3.28), and 2.94 could be preliminarily used as the upper limit of the critical value for early lung cancer in males. For males in the middle-late lung cancer group, the fibrinogen level distribution was greater (2.27–5.80) and concentrated in category 7 (4.29–5.80), and 3.91 could serve as the lower limit of middle-late lung cancer in males. For female patients, both t early stage and the middle-late stage of lung cancer have relatively smaller fibrinogen level distributions (1.26–3.78) that are mostly distributed among lower categories, particularly category 2 (2.27–2.77), and 3.91 could be used as the upper limit of the critical value for the early stage of lung cancer.

Previous investigators reported that patients with higher plasma fibrinogen levels tend to have poor progression-free survival (PFS) and overall survival (OS) (15,16). Therefore, fibrinogen may serve as a candidate biomarker for disease monitoring and prognostic evaluation in patients with lung cancer. In future studies, we will attempt to obtain follow-up data from patients for further analysis.

Real-world data mining for scientific creation

In clinical medicine research, the discovery of research

hypotheses is classically a difficult task. Among traditional methods for establishing medical research hypotheses, scientific creations typically arise from the experience summarized by medical scientists and originate from the study of the existing scientific literature. Researchers then focus on project design, data collection, clinical trial implementation, evidence-based analysis according to these hypotheses, and finally verification of these results using different experimental and statistical methods. Although these hypotheses are based on professional medical knowledge and evidence that has been revealed by scientific research, they may not be supported by real-world data, leading to the failure of these research projects.

Real-world research is a relatively new method for exploring scientific hypotheses in the medical field and can involve the implementation of machine learning and data mining technology. By applying these methods to real-world data that are collected from all types of clinical areas and further applying statistical methods to verify the mined results, we can actively explore the features and knowledge hidden in the data and then establish reliable scientific hypotheses. This process often leads to a solid foundation for scientific research and therefore greatly improves the success of research. This paper is a good example of real-world data mining.

Some limitations in this study should be addressed. Our data only included patients with lung cancer from a single center, and the sample size of patients with middle-late lung cancer was relatively small. Although it is difficult to avoid bias in a single-center retrospective study, we tried to reduce bias using the following methods: strictly controlling the inclusion or exclusion criteria of the subjects to reduce selection bias; using objective data to reduce information bias in subsequent data analysis as much as possible; multivariate analysis and artificial elimination of known confounding effects to reduce confounding bias; and setting a validation cohort to validate the model to avoid bias caused by overfitting in data mining. Further studies will be performed to continue exploring the distribution of the characteristics of fibrinogen and sex in healthy or high-risk populations and different stages of lung cancer (I, II, III, IV) based on multicenter data research, which would provide new clues and improved research support for the diagnosis of lung cancer.

Conclusions

In this paper, we integrated different machine learning methods, including the MDL variable selection model, NB classification model, KM clustering model, NMF feature extraction model, and DT explicit expression model, to systematically explore the distribution of the characteristics of different fibrinogen levels among clinical stages of lung cancer between the sexes. First, the MDL algorithm was applied to select the relevant variables for distinguishing the target variable. Then, a supervised NB classification algorithm was applied to establish a classification model, and its performance was evaluated. The reliability of the model was validated using a validation cohort. Then, an unsupervised KM algorithm was used to explore the distribution of the characteristics. Next, an NMF algorithm was applied to extract the features from the data. Finally, a supervised DT algorithm was proposed. The results of this paper demonstrated that the comprehensive application of a variety of supervised and unsupervised machine learning algorithms could effectively explore the knowledge and laws contained in large samples of real-world clinical data.

Compared to traditional clinical scientific research, machine learning based on real-world data can better promote the development of scientific research, especially creating a scientific hypothesis. Therefore, research periods can be significantly reduced, scientific research output can be improved quantitatively, and the breadth of the research can be increased qualitatively by expanding the train of thought of research and identifying a new way of thinking based on different aspects.

Acknowledgments

The authors wish to thank Mr. Guozhang Dong for his crucial help in checking the staging data of lung cancer cases.

Funding: This work was supported by Six Summit Investigator Grant of Jiangsu Province (WSW-014), Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD2018-87), National Science Foundation of China (81872378, 81672295, 81572261, and 81501977), China Postdoctoral Science Foundation (2018M640465), the Project of Jiangsu Provincial Medical Talent (ZDRCA2016033) and the Key Project of Nantong Livelihood Science and Technology (MS22018013).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at http://dx.doi.

org/10.21037/atm-20-4704

Data Sharing Statement: Available at http://dx.doi.org/10.21037/atm-20-4704

Peer Review File: Available at http://dx.doi.org/10.21037/atm-20-4704

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-20-4704). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted according in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the ethics committee of Jiangsu Cancer Hospital (IRB approval: 2017-K-001). As a retrospective study without any effects on individual patient outcomes, the ethical review board waived the need for individual patient informed consent.

References

1. Hoffman RM, Sanchez R. Lung Cancer Screening. Med Clin North Am 2017;101:769-85.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018;68:7-30.
3. Payne S. 'Smoke like a man, die like a man'?: a review of the relationship between gender, sex and lung cancer. Soc Sci Med 2001;53:1067-80.
4. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13:1.
5. Thunnissen E, van der Oord K, den Bakker M. Prognostic and predictive biomarkers in lung cancer. A review. Virchows Arch 2014;464:347-58.
6. de Moerloose P, Casini A, Neerman-Arbez M. Congenital Fibrinogen Disorders: An Update. Semin Thromb Hemost 2013;39:585-95.
7. Wang Y, Fuller GM. The putative role of fibrin fragments in the biosynthesis of fibrinogen by hepatoma cells. Biochem Biophys Res Commun 1991;175:562-7.
8. Buccheri G, Ferrigno D, Ginardi C, et al. Haemostatic abnormalities in lung cancer: prognostic implications. Eur J Cancer 1997;33:50-5.
9. Liu X, Shi B. Progress in research on the role of fibrinogen in lung cancer. Open Life Sci 2020;15:326-30.
10. Key NS, Khorana AA, Kuderer NM, et al. Venous Thromboembolism Prophylaxis and Treatment in Patients With Cancer: ASCO Clinical Practice Guideline Update. J Clin Oncol 2020;38:496-520.
11. Rachidi S, Metelli A, Riesenberg B, et al. Platelets subvert T cell immunity against cancer via GARP-TGFβ axis. Sci Immunol 2017;2:eaai7911.
12. Placke T, Salih HR, Kopp HG. GITR Ligand Provided by Thrombopoietic Cells Inhibits NK Cell Antitumor Activity. J Immunol 2012;189:154-60.
13. Mammadova-Bach E, Zigrino P, Brucker C, et al. Platelet integrin α6β1 controls lung metastasis through direct binding to cancer cell-derived ADAM9. JCI Insight 2016;1:e88245.
14. Volz J, Mammadova-Bach E, Gil-Pulido J, et al. Inhibition of platelet GPVI induces intratumor hemorrhage and increases efficacy of chemotherapy in mice. Blood 2019;133:2696-706.
15. Sheng L, Luo M, Sun X, et al. Serum fibrinogen is an independent prognostic factor in operable nonsmall cell lung cancer. Int J Cancer 2013;133:2720-5.
16. Zhu LR, Li J, Chen P, et al. Clinical significance of plasma fibrinogen and D-dimer in predicting the chemotherapy efficacy and prognosis for small cell lung cancer patients. Clin Transl Oncol 2016;18:178-88.

Cite this article as: Yin F, Zhu H, Hong S, Sun C, Wang J, Sun M, Xu L, Wang X, Yin R. An application of machine learning based on real-world data: Mining features of fibrinogen in clinical stages of lung cancer between sexes. Ann Transl Med 2021;9(8):623. doi: 10.21037/atm-20-4704