



Article

Determining Risk Factors Associated with Depression and Anxiety in Young Lung Cancer Patients: A Novel Optimization Algorithm

Yu-Wei Fang ^{1,2}  and Chieh-Yu Liu ^{3,4,*} 

¹ Department of Nephrology, Shin Kong Memorial Wu Ho-Su Hospital, Taipei 111, Taiwan; M005916@ms.skh.org.tw

² Department of Medicine, Fu-Jen Catholic University, New Taipei 242, Taiwan

³ Biostatistical Consulting Lab, Department of Speech Language Pathology and Audiology, National Taipei University of Nursing and Health Sciences, Taipei 112, Taiwan

⁴ Department of Teaching and Research, Taipei City Hospital, Taipei 106, Taiwan

* Correspondence: chiehyu@ntunhs.edu.tw; Tel.: +886-2-28227101 (ext. 6205/3312)

Abstract: *Background and Objectives:* Identifying risk factors associated with psychiatrist-confirmed anxiety and depression among young lung cancer patients is very difficult because the incidence and prevalence rates are obviously lower than in middle-aged or elderly patients. Due to the nature of these rare events, logistic regression may not successfully identify risk factors. Therefore, this study aimed to propose a novel algorithm for solving this problem. *Materials and Methods:* A total of 1022 young lung cancer patients (aged 20–39 years) were selected from the National Health Insurance Research Database in Taiwan. A novel algorithm that incorporated a *k*-means clustering method with *v*-fold cross-validation into multiple correspondence analyses was proposed to optimally determine the risk factors associated with the depression and anxiety of young lung cancer patients. *Results:* Five clusters were optimally determined by the novel algorithm proposed in this study. *Conclusions:* The novel Multiple Correspondence Analysis-*k*-means (MCA-*k*-means) clustering algorithm in this study successfully identified risk factors associated with anxiety and depression, which are considered rare events in young patients with lung cancer. The clinical implications of this study suggest that psychiatrists need to be involved at the early stage of initial diagnose with lung cancer for young patients and provide adequate prescriptions of antipsychotic medications for young patients with lung cancer.

Keywords: young lung cancer; depression; anxiety; multiple correspondence analysis; *k*-means clustering



Citation: Fang, Y.-W.; Liu, C.-Y. Determining Risk Factors Associated with Depression and Anxiety in Young Lung Cancer Patients: A Novel Optimization Algorithm. *Medicina* **2021**, *57*, 340. <https://doi.org/10.3390/medicina57040340>

Academic Editor: Camelia Diaconu

Received: 19 February 2021

Accepted: 24 March 2021

Published: 1 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Lung cancer is a very aggressive malignant disease; people who smoke or are exposed to polluted environments or with genetic mutation may be at significantly higher risk of lung cancer [1–3]. Published studies have shown that Asian women who never smoke still have a higher risk of lung cancer compared with women in European countries or the United States [4,5]. Based on a worldwide report issued by the International Agency for Research on Cancer in 2018, for both males and females, lung cancer had become the most prevalent cancer globally (with incidence rate of 11.6% of all cancers) and had been ranked as the leading cause of cancer death (death rate of 18.4% of the total cancer deaths). The economic burden of treatments and care for lung cancer has also globally increased in recent years [6,7]. Published studies have showed that lung cancer is significantly associated with older age (70 years old being the average age of initial diagnosis) [7,8], but very low incidence rate in young people aged 20–40 years around the world [8]. In recent decades, due to the dramatic improvements of clinical treatments and screening techniques for lung cancer, the survival of newly diagnosed lung cancer patients has been

significantly prolonged [9,10] and the incidence rate in young patients with lung cancer also increased [1,5,8]. Published studies also showed that young patients with lung cancer had better treatment outcomes of receiving surgery, chemotherapy, or radiotherapy and have relatively longer relapse-free survival, which indicates that young lung cancer patients are more likely to have prolonged survival [11–13].

Therefore, young patients with lung cancer are a noteworthy group of patients, because they have obviously lower incidence rate and prevalence rate than middle-age or elder people and they may have longer survival time. Liu et al. (2019) [14] used a retrospective review of patients with lung cancer in one hospital in China from January 2010 to June 2017, the prevalence of lung cancer in young adults aged between 18 and 35 years old was 1.37%; and Rich et al. (2015) [15] also used a retrospective cohort review using a validated national audit dataset and the results showed that the prevalence of lung cancer in young adults aged between 18 and 39 years was 0.5%. The overall incidence and prevalence in elder age groups (>50 years old) was increasing in recent decade, however, the incidence and prevalence of lung cancer in young adults (<40 years old) can be still regarded as relatively low in nowadays global cancer epidemiology. Recent studies showed that young lung cancer survivors are also at a high risk of psychiatric diseases, such as anxiety and depression in the following years of survival [16–18]. However, most of the published studies used self-reported scales or questionnaires to measure anxiety and depression instead of using diagnoses by psychiatrists; therefore, the so-called depression or anxiety in published studies can solely regarded as depression symptoms or anxiety symptoms. For example, Yan et al. [17] showed that the anxiety and depression prevalence rates of lung cancer patients were 43.5% and 57.1% by using the Hospital Anxiety and Depression Scale (HADS), which look high proportions in lung cancer patients. In addition, if young lung cancer patients who may have prolonged survival and are at high risk of psychiatrist-confirmed depression and anxiety, they will consume considerable medical resources due to the additional treatments for psychiatric diseases [6,19]. Nevertheless, there is still a lack of literature investigating risk factors associated with psychiatrist-confirmed depression and anxiety in young lung cancer patients. This study was aimed to develop a novel algorithm for identifying risk factors for psychiatrist-confirmed anxiety and depression in young lung cancer patients aged 20–39 years old by using the population-based database (National Health Insurance Research Database (NHIRD) in Taiwan), which can assist clinicians or young patients with lung cancer in preventing anxiety and depression at early stages.

2. Materials and Methods

2.1. Study Design and Study Database

This study design of this research adopted the secondary analysis of longitudinal data from NHIRD. The study database used here was retrieved from the NHIRD in Taiwan. Since the National Health Insurance (NHI) program was launched on 1 March 1995, the NHI program provided healthcare service coverage to more than 99% of the population by 2017 [20]. The NHIRD includes medical reimbursement records for outpatient and inpatient healthcare services, hospital or clinic visits, dental service visits and traditional Chinese medicine service visits. All of the reimbursement records for diagnostic and medical-related procedures for diseases are based on the international classification of diseases (ICD)—ninth and tenth revisions (after 1 January 2016 [21]) of the clinical modification (CM, or ICD-9-CM and ICD-10-CM, respectively)—and on a procedure coding system for all medical service claims.

2.2. Ethics Statement

The ethical review of this study was approved by the Institutional Review Board of the School of Nursing, National Taipei University of Nursing and Health Sciences (approval number: IRB# CN-IRB-2011-063). The date of approval was 23 October 2011. The encryption and protection of the personal information from the NHIRD were performed by the National Health Insurance Administration in Taiwan by using a complex double

encryption procedure. In addition, because the present study was a secondary data analysis, written informed consent forms were not required from the recruited or selected patients. This study was also registered at Open Science Framework (OSF, reference osf.io/fkxm8 (accessed on 15 March 2021)).

2.3. Study Population and Possible Risk Factors Selection

The ICD-9-CM codes that were used to define patients with depression were 296.2X–296.3X, 300.4 and 311.X and the ICD-9-CM codes used to define patients with anxiety were 300.XX, 291.89 and 292.89. In Taiwan, if cancer patients are suspected of having depression or anxiety, they are referred by the oncologists to psychiatrists, which is recorded as the first National Health Insurance (NHI) outpatient visit. After the referral, the cancer patients receive some psychological tests by clinical psychologists and the cancer patients are diagnosed by psychiatrists again to determine if they need anti-depressant or anti-anxiety medications; this is recorded as the second NHI psychiatric visit. After a period of time, the cancer patients need to be confirmed again by psychiatrists; therefore, to confirm that a cancer patient has depression or anxiety usually needs at least three outpatient visits and the prescription of anti-depressant or anti-anxiety drugs. In this study, young lung cancer patients that were aged 20–39 years and who were newly diagnosed with lung cancer (ICD-9-CM code = 162.XX) between 1 January 2001, and 31 December 2007, were retrieved from the NHIRD. Young lung cancer patients who died or withdrew from the NHI program during the study period were excluded. Young patients with lung cancer who had been diagnosed with baseline psychiatric diseases, such as depressive disorder (ICD-9-CM codes: 296.2X–296.3X, 300.4 and 311.X), anxiety states (ICD-9-CM codes: 300.XX, 291.89 and 292.89), bipolar disorders (ICD-9-CM codes: 296.0, 296.1, 296.4, 296.5, 296.6, 296.7, 296.8, 296.80 and 296.89), or alcohol-induced mental disorders (ICD-9-CM codes: V113, 9800, 2650, 2651, 3575, 4255, 3050, 291, 303 and 571.0–571.3) between 1 January and 31 December in 2001 were also excluded. In order to avoid selecting false-positive patients with depression and anxiety, young lung cancer patients with at least three consecutive corresponding diagnoses were eligible to be coded as having depression and anxiety.

The possible risk factors associated with depression and anxiety among lung cancer patient were determined based on Park et al. [19], who investigated if hypertension, diabetes mellitus, history of tuberculosis, liver disease (liver cancer and liver cirrhosis), end-stage renal disease, coronary artery disease (including heart failure), stroke (ischemic stroke and hemorrhage stroke) and Chronic obstructive pulmonary disease (COPD) are risk factors associated with anxiety and depression after surgical treatment for lung cancer; and Clarke and Currie [20], who took into account heart disease, stroke, cancer, diabetes mellitus, rheumatoid arthritis and asthma as the possible risk factors associated with depression and anxiety in cancer patients. Therefore, in this study, we took into account diabetes mellitus (DM), hypertension, asthma, liver cirrhosis, COPD, autoimmune diseases (including rheumatoid arthritis, systemic lupus erythematosus and aplastic anemia), cerebral diseases (including ischemic stroke, hemorrhage stroke and transient ischemic attack (TIA)), heart failure, hepatitis B virus (HBV), renal diseases and osteoporosis.

2.4. Combining Multiple Correspondence Analysis and the K-Means Clustering Algorithm with *v*-Fold Cross-Validation (MCA-*k*-Means Clustering Algorithm)

The raw data matrix was first transformed into a matrix with solely index variables (i.e., encoded as 0 or 1) through multiple correspondence analysis (MCA) [21,22], which was the data preprocessing procedure for the raw data matrix. The index variables indicate the levels of all of the categorical variables in this study. The MCA then converted all index variables into multi-dimensional Euclidean coordinates. The multi-dimensional Euclidean coordinate matrix derived from the MCA could be considered a high-dimensional dataset that could be carried into the further optimal clustering algorithm. In order to determine the optimal clustering in the high-dimensional dataset obtained from the MCA, the *k*-means clustering algorithm with *v*-fold cross-validation was applied to obtain the optimal clustering. The algorithm is described in detailed in the following:

2.4.1. Step 1. Multiple Correspondence Analysis

Let $M_{I \times K}$ be the raw data matrix with I subjects and k categorical variables.

(1) Transform the raw data matrix into a Burt matrix:

- If a categorical variable is binary, then place it in the Burt matrix as an original variable matrix.
- If a categorical variable has more than two levels (i.e., $J_k > 2$ levels), then convert this variable into an index variable (containing only 0 and 1); this forms an indicator matrix $I \times J_k$ where each column contains index variables coded with 0 or 1.
- Place all index variable columns together to form the indicator matrix $X_{I \times J}$.
- Calculate the Burt matrix as $(X_{I \times J})' \cdot X_{I \times J}$.

(2) Calculate the column and row coordinates as follows:

- The total orders of $M_{I \times K}$ (N) are observed and the probability matrix is defined as $P = N^{-1}X$.
- Define r as the vector of the row totals of P (i.e., $r = P1$, where 1 is a unit vector of ones) and define c as the vector of the column totals of P . Then, $D_c = \text{diag}\{c\}$ and $D_r = \text{diag}\{r\}$.
- Calculate the Euclidean coordinates by using a singular value decomposition method as follows:

$$D_r^{-\frac{1}{2}} \left(Z - rc^T \right) D_c^{-\frac{1}{2}} = P \Delta Q^T$$

where Δ and $\Lambda = \Delta^2$ are the diagonal matrix of singular values and the matrix containing eigenvalues, respectively. Therefore, the row and column coordinate matrices (F and G , respectively) are calculated as follows:

$$F = D_r^{-\frac{1}{2}} P \Delta$$

$$G = D_c^{-\frac{1}{2}} Q \Delta$$

(3) The number of dimensions is determined using an inertia value as follows:

- The inertia value is calculated based on a Pearson chi-squared (χ^2) value from the rows and columns to identify their coordinate centers as follows:

$$d_r = \text{diag}\{FF^T\} \text{ and } d_c = \text{diag}\{GG^T\}.$$

- If a subset of F or G is selected, then the inertia values for the row and column coordinates are calculated as:

$$Inertia_r = \frac{\text{diag}\{FF'^T\}}{N} \text{ and } Inertia_c = \frac{\text{diag}\{GG'^T\}}{N},$$

where F' and G' are subsets of F and G .

2.4.2. Step 2. K-Means Clustering with v-Fold Cross-Validation

The k -means clustering algorithm with v -fold cross-validation was applied to analyze the F and G that were obtained from the MCA [23,24]. The algorithm is as follows:

- (1) Determination of the range of numbers of clusters for the k -means clustering algorithm: In this study, the number was set from $k = 2$ to n , where $n \leq 10$;
- (2) Determination of the initial cluster centers: The initial cluster centers were selected at random;
- (3) Iteration scheme: Assigning all index variables to their nearest cluster centers. The Euclidean distance was used as the distance measurement in the iterative classification scheme;

- (4) To determine the optimal clustering, *v*-fold cross-validation was applied to estimate the optimal number of clusters and the optimal clustering. The details of the *v*-fold cross-validation are as follows:
 - (a) Divide **F** or **G** into *v* folds (denoted F_i or G_i , $i = 1, \dots, v$), in this study, we set $v = 5$;
 - (b) For $i = 1$ to v , take F_i or G_i as the testing set and $\{F\} \setminus F_i$ or $\{G\} \setminus G_i$ as the training sets;
 - (c) Compute the mean Euclidean distances, which are called the clustering costs in this study, within each cluster of training sets, set these as the new cluster centers and replace the cluster centers of the previous step;
 - (d) Compute the mean Euclidean distances of each index variable (or the level of all of the categorical variables) of the testing set from the new cluster centers derived from the training sets;
- (5) Iterate from (1);
- (6) If $k = j$, which indicates the minimum mean Euclidean distances (i.e., minimal clustering cost) of each index variable of the testing set, j would be the optimum number of clusters.
- (7) Clustering stopping rule: If $|\bar{D}_{j+1} - \bar{D}_j| < 0.01$, then stop further dividing and clustering.
- (8) Regarding the determination of number of clusters, we adopted the method proposed by Wang [25], the optimal algorithm will iterate in order to classify factors into different numbers of clusters, calculate the cluster cost (in this study, we used the mean sum of squares within clusters as the cluster cost measurement) and compare the sums of squares between clusters. If the sum of squares of k clusters did not show statistically significant difference from $k + 1$ clusters, the optimal number of clusters is determined as k .

The MY Structured Query Language (MySQL) was used for selection, linkage, processing and cleaning of the dataset from the NHIRD. The algorithm we proposed in this study was implemented with STATISTICA Data Miner ver. 10.0 (StatSoft, Inc., Tulsa, OK, USA).

3. Results

In the present study, 1022 young lung cancer patients aged 20–39 years were studied and their demographic information is shown in Table 1. The study sample comprised 520 male (50.9%) and 502 female patients (49.1%); 154 of the patients were aged 20–29 years old (15.1%) and 868 patients were aged 30–39 years old (84.9%).

Table 1. Demographic information of the study sample ($n = 1022$).

Variable	<i>n</i>	(%)
Sex		
Female	502	49.1
Male	520	50.9
Age		
20–29 y	154	15.1
30–39 y	868	84.9
Charlson comorbidity index (CCI)		
CCI = 0	870	85.1
CCI = 1	91	8.9
CCI ≥ 2	61	6
Diabetes mellitus (DM)		
Yes	23	2.3
No	999	97.7

Table 1. *Cont.*

Variable	<i>n</i>	(%)
Hypertension		
Yes	23	2.3
No	999	97.7
Asthma		
Yes	16	1.6
No	1006	98.4
Liver cirrhosis		
Yes	9	0.9
No	1013	99.1
Chronic obstructive pulmonary disease (COPD)		
Yes	51	5
No	971	95
Autoimmune diseases		
Yes	8	0.8
No	1014	99.2
Cerebral diseases		
Yes	11	1.1
No	1011	98.9
Heart failure		
Yes	2	0.2
No	1020	99.8
Hepatitis B virus (HBV)		
Yes	34	3.3
No	988	96.7
Renal diseases		
Yes	6	0.6
No	1016	99.4
Osteoporosis		
Yes	16	1.6
No	1006	98.4
Depression		
Yes	25	2.4
No	997	97.6
Anxiety		
Yes	15	1.5
No	1007	98.5

As a result of the k-means clustering of **F** and **G**, which were Euclidean coordinate matrixes derived from the multiple correspondence analysis (MCA) and by using *v*-fold cross-validation, the clustering costs of different numbers used for the k-means clustering algorithm are shown in Figure 1. According to the results shown in Figure 1, on the basis of the clustering cost, there was no statistically significant difference between using five clusters or six clusters. Based on the principal of parsimony of clustering, the optimum number of clusters was determined to be five. Table 2 presents the clustering results that comprise these five clusters. Table 2 indicates that anxiety was clustered with osteoporosis and depression was clustered with the lack of diabetes mellitus (DM), Charlson comorbidity index (CCI) = 0 and female sex.

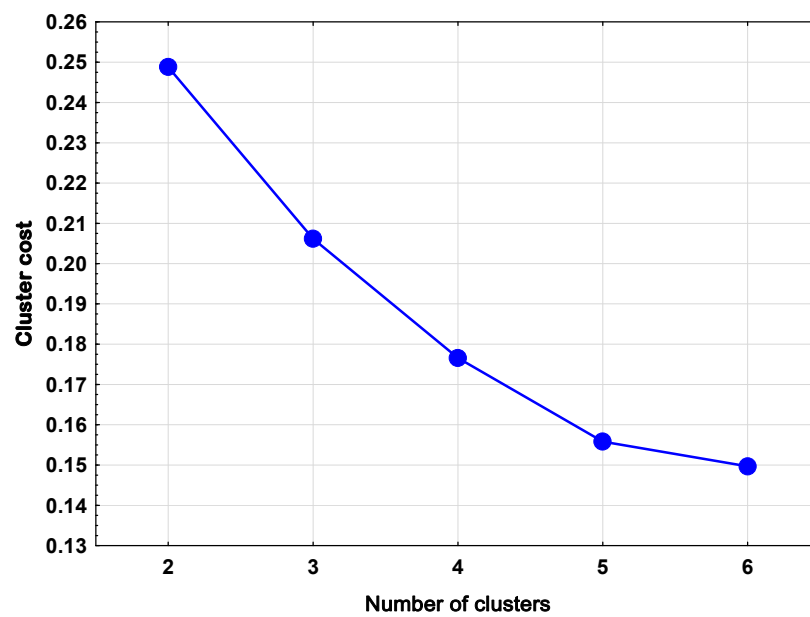


Figure 1. Cluster costs of different numbers of clusters resulting from k-means clustering combined with *v*-fold cross-validation.

Table 2. Results of the multiple correspondence analysis (MCA) and k-means algorithm with *v*-fold cross-validation.

Variable	Final Classification
Autoimmune disease = Yes	1
Cerebral disease = Yes	1
Heart failure = Yes	1
Osteoporosis = Yes	2
Anxiety = Yes	2
Depression = Yes	3
DM = No	3
Age: 20–29 y	3
Age: 30–39 y	3
CCI = 0	3
Sex = Female	3
DM = Yes	4
Hypertension = Yes	4
Asthma = Yes	4
Liver cirrhosis = Yes	4
COPD = Yes	4
HBV = Yes	4
CCI ≥ 2	4
Depression = No	5
Hypertension = No	5
Asthma = No	5
Liver cirrhosis = No	5
COPD = No	5
Autoimmune disease = No	5
Cerebral disease = No	5
Heart failure = No	5
HBV = No	5
Osteoporosis = No	5
Anxiety = No	5
CCI = 1	5
Sex = Male	5

Note: DM = Diabetes mellitus; CCI = Charlson comorbidity index; COPD = Chronic obstructive pulmonary disease; HBV = Hepatitis B virus.

In addition, in the present study, a control arm statistical analysis was also performed using a multiple logistic regression model, which is the most widely used method for investigating risk factors associated with diseases. Table 3a shows the results of score tests of both dependent variables—depression and anxiety—for each independent variable. In Table 3a, no statistical significance was observed for any of the independent variables with these two dependent variables, indicating that using a stepwise variable selection strategy (forward or backward variable selection) cannot be used to find any statistically significant predictors. Furthermore, Table 3b shows the results of the multiple logistic regression model (without variable selection procedures), which also indicated that there were no statistically significant predictors (except for constant terms for both dependent variables).

Table 3. (a) Score test results for each variable of the logistic regression model. (b) Results of multiple logistic regression models for depression and anxiety.

(a)					
Variable	DV = Depression		DV = Anxiety		
	Score	<i>p</i> -Value	Score	<i>p</i> -Value	
Sex: Male vs. Female	0.485	0.486	0.108	0.742	
Age: 30–39 vs. 20–29 years	0.017	0.895	0.036	0.85	
CCI = 1 vs. CCI = 0	2.505	0.113	0.368	0.544	
CCI ≥ 2 vs. CCI = 0	1.661	0.197	0.966	0.326	
DM: Yes vs. No	0.59	0.442	0.35	0.554	
Hypertension: Yes vs. No	0.357	0.55	0.35	0.554	
Asthma: Yes vs. No	0.408	0.523	2.571	0.109	
Liver cirrhosis: Yes vs. No	0.228	0.633	0.135	0.713	
COPD: Yes vs. No	0.053	0.818	0.09	0.764	
Autoimmune: Yes vs. No	0.202	0.653	0.12	0.729	
Cerebral diseases: Yes vs. No	0.279	0.597	0.166	0.684	
Heart failure: Yes vs. No	0.05	0.823	0.03	0.863	
HBV: Yes vs. No	1.74	0.187	0.524	0.469	
Renal diseases: Yes vs. No	0.151	0.697	0.09	0.764	
Osteoporosis: Yes vs. No	0.408	0.523	2.571	0.109	

(b)								
Variable	DV = Depression				DV = Anxiety			
	Beta	S.E.	Odds Ratio (OR)	<i>p</i> -value	Beta	S.E.	Odds Ratio (OR)	<i>p</i> -value
Sex: Male vs. Female	−0.352	0.418	0.703	0.399	−0.122	0.530	0.885	0.818
Age: 30–39 vs. 20–29 years	−0.010	0.561	0.990	0.986	0.038	0.781	1.038	0.961
CCI = 1 vs. CCI = 0	−17.329	4055.844	<0.001	0.997	0.302	0.872	1.353	0.729
CCI ≥ 2 vs. CCI = 0	1.377	0.829	3.964	0.097	−15.014	4327.707	<0.001	0.997
DM: Yes vs. No	−18.980	7844.428	<0.001	0.998	−14.048	6665.142	<0.001	0.998
Hypertension: Yes vs. No	1.217	1.092	3.377	0.265	−16.095	7460.922	<0.001	0.998
Asthma: Yes vs. No	−17.069	8269.606	<0.001	0.998	18.338	6580.883	92,044,936.212	0.998
Liver cirrhosis: Yes vs. No	−16.960	11,705.571	<0.001	0.999	−16.057	11,754.148	<0.001	0.999
COPD: Yes vs. No	−0.177	1.116	0.838	0.874	−16.910	6580.883	<0.001	0.998
Autoimmune: Yes vs. No	−17.528	12,892.859	<0.001	0.999	−17.150	13,490.401	<0.001	0.999
Cerebral diseases: Yes vs. No	−17.579	11,033.665	<0.001	0.999	−16.307	11,052.028	<0.001	0.999
Heart failure: Yes vs. No	−18.191	25,475.907	<0.001	0.999	−16.436	26,494.679	<0.001	1.000
HBV: Yes vs. No	0.225	0.962	1.252	0.815	−16.248	6243.383	<0.001	0.998
Renal diseases: Yes vs. No	−18.808	16,186.569	<0.001	0.999	−14.876	14,129.940	<0.001	0.999
Osteoporosis: Yes vs. No	−17.344	9805.003	<0.001	0.999	1.507	1.131	4.511	0.183
Constant	−3.468	0.561	0.031	<0.001	−4.152	0.778	0.016	<0.001

Note: S.E. = Standard Error; DM = Diabetes mellitus; CCI = Charlson comorbidity index; COPD = Chronic obstructive pulmonary disease; HBV = Hepatitis B virus.

4. Discussion

The objective of this study aimed to develop a novel algorithm for identifying risk factors for anxiety and depression in young lung cancer patients aged 20–39 years by using the population-based database (National Health Insurance Research Database (NHIRD) in Taiwan), which are regarded rare events and very limited number of methods were proposed to solve this problem. A novel algorithm was proposed in this study which

integrated v -fold cross-validation into MCA- k -means clustering for solving the problem of determining risk factors associated with rare events.

Compared with the results of a univariate analysis using traditional multiple logistic regression analysis, which is a widely used method for determining risk factors associated with diseases (see Table 3), the results showed that none of the risk factors were statistically significantly associated with anxiety and depression, respectively, in young patients with lung cancer. Moreover, some parameter estimates were very unreliable because of their large standard errors (even bigger than the parameter estimates). In Table 3a, for the depression outcome variable, CCI = 1 vs. CCI = 0, DM, asthma, liver cirrhosis, autoimmune diseases, cerebral diseases, heart failure, renal diseases and osteoporosis indicated that parameter estimates were unreliable and exhibited extremely low odds ratios (ORs); for the anxiety outcome variable, CCI ≥ 2 vs. CCI = 0, DM, hypertension, asthma, liver cirrhosis, chronic obstructive pulmonary disease (COPD), autoimmune diseases, cerebral diseases, heart failure, hepatitis B (HBV) and renal diseases also indicated that the parameter estimates were unreliable and exhibited extremely low odds ratios (ORs), or an extremely high OR for asthma. Previous studies have indicated that parameter estimation methods such as maximum likelihood estimation provide biased or inestimable estimates for rare events [26,27]. According to King and Zeng (2001) [28], logistic regression would sharply underestimate the probability of rare events. For resolving the problems, some methods have been proposed, but there is still a lack of optimal methods and agreements on how to better estimate the coefficient of logistic regression for rare event data. In this study, not only were the dependent variables (depression and anxiety) rare events, but so were the independent variables, which may have resulted in many zeros in the database and the estimation of the standard error may have been biased. The novel algorithm proposed in this study can be considered to be a good approach for resolving rare event problems. In addition, compared with the results using self-reported questionnaire or inventory, such as Yan et al. [17], which used binary logistic regression analysis and the results showed that the risk factors of both anxiety and depression were lack of surgery and age; however, binary logistic regression did not successfully identify statistically significant risk factors in this study and the difference can be resulted from different operational definitions of depression and anxiety. Both kinds of studies using self-reported questionnaires or ICD-9-CM codes by psychiatrist-confirmed diagnoses provide different contributions to the clinical practices. Studies using self-reported questionnaires or inventory to measure depression and anxiety are more likely to look for factors associated with the self-perceived depression symptom and anxiety symptoms, which may be easier to express by patients themselves and some behavior interventions may be suggested, such as exercise, focus group consultant or health promotion life adjustment. However, the results of the current study using ICD-9-CM codes of depression and anxiety which are confirmed by psychiatrists, what young patients with lung cancer need are not only behavior interventions, but also the prescriptions of antidepressant drugs or anti-anxiety drugs, or the psychiatric hospitalization.

The advantages of the MCA- k -means clustering algorithm proposed in this study are: (1) the adoption of the clustering-based method to determine risk factors associated with rare events, which may avoid the parameter estimation problems encountered when using conventional logistic regression models; (2) the algorithm can take more than one dependent variable (≥ 2) into account simultaneously, especially for easily confused diseases, for example, anxiety and depression in this study. In comparison with a logistic regression model, it deals with only one dependent variable at a time. (3) The algorithm determines the optimum number of clusters by using the v -fold cross-validation algorithm; through the repeated random sub-sampling scheme, all observations were used for both the training and validation sets and each observation was used for validation exactly once, which can help determine the optimum number of clusters with less influence from rare event data, such as the dataset used in this study.

Regarding the final clustering results of this study (see Table 2), the results indicated that anxiety was clustered with osteoporosis and depression was clustered with the lack

of DM, CCI = 0 and female sex in young patients with lung cancer. These factors were optimally clustered with anxiety and depression. The results obtained in this study are validated by other studies that have indicated that patients with anxiety and osteoporosis easily encounter more complications than those with several other disease groups [29–31]. The results of this study indicate that young patients with lung cancer and osteoporosis are also at a high risk for the onset of anxiety. In addition, young female lung cancer patients were also at a higher risk of the onset of depression. Previously published studies have shown that female cancer patients are at significantly higher risk of depression than males [29,32,33]. In this study, the clustering results also supported that young female lung cancer patients were at a higher risk of the onset of depression.

This study still had some limitations. First, although the National Health Insurance (NHI) program in Taiwan covers more than 98% of the Taiwanese population [34–36], the NHIRD does not provide information about some potential confounding factors, such as smoking, alcohol consumption, exercise habits, diet and lifestyle, which may also influence the association with the risk of anxiety and depression. Second, some young lung cancer patients who experience anxiety and depression may not consult psychiatrists; they usually express their concerns about their cancer diseases to their oncologists and the oncologists may easily neglect or ignore their patients' anxiety and depression symptoms. Thus, cancer patients may search for religious help or may isolate themselves from people or medical professionals; therefore, the number of patients with anxiety and depression may be underestimated. Third, because the young patients with lung cancer enrolled in this study were primarily of the Chinese or Han ethnicities, the results derived from the novel algorithm proposed here require further examination and validation for generalization to other ethnicities. Furthermore, according to Lu et al. (2019) [37], in recent decades, the overall incidence of lung cancer initially increased and then gradually decreased. The surgical rate and radiotherapy rate for lung cancer showed a general downward trend, while the chemotherapy rate experienced a significantly increasing trend [30]. Although the five-year relative survival rate has increased over the years, it has remained very low for the last 20 years [31]. Therefore, this study, which used a nationwide database from 2001 to 2007, can still provide useful findings for clinicians.

5. Conclusions

The novel MCA–k-means clustering algorithm in this study successfully identified risk factors associated with anxiety and depression, which are considered rare events in young patients with lung cancer. The clinical implications of this study suggest that psychiatrists need to be involved at the early stage of initial diagnose with lung cancer for young patients and provide adequate prescriptions of antipsychotic medications for young patients with lung cancer.

Author Contributions: Drafting of the article: Y.-W.F.; critical revision of the article for important intellectual content: Y.-W.F. and C.-Y.L.; final approval of the article: C.-Y.L.; statistical expertise: C.-Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by an industry–academia collaboration grant whose grant number is DSLPA-PC-107-003.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and it was approved by the Institutional Review Board of the School of Nursing, National Taipei University of Nursing and Health Sciences (approval number: IRB# CN-IRB-2011-063).

Informed Consent Statement: Patient consent was waived because the encryption and protection of the personal information from the NHIRD were performed by the National Health Insurance Administration in Taiwan by using a complex double-encryption procedure. As this present study was a secondary data analysis, written informed consent forms were not required from the recruited or selected patients.

Data Availability Statement: The study dataset (NHIRD) was not publicly archived; to access it, an application from the Bureau of National Health Insurance in Taiwan is needed. The application website is: <https://www.nhi.gov.tw> (the access date was 10 December 2012).

Acknowledgments: The authors of this study are very grateful to the National Health Insurance Administration for providing the National Health Insurance claim database and to the Health Data Value-Added Center of the Ministry of Health and Welfare of Taiwan for maintaining the National Health Insurance Research Database (NHIRD).

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

- Cheng, T.-Y.D.; Cramb, S.M.; Baade, P.D.; Youlden, D.R.; Nwogu, C.; Reid, M.E. The international epidemiology of lung cancer: Latest trends, disparities, and tumor characteristics. *J. Thorac. Oncol.* **2016**, *11*, 1653–1671. [[CrossRef](#)] [[PubMed](#)]
- Cho, J.H.; Zhou, W.; Choi, Y.-L.; Sun, J.-M.; Choi, H.; Kim, T.-E.; Dolled-Filhart, M.; Emancipator, K.; Rutkowski, M.A.; Kim, J. Retrospective molecular epidemiology study of PD-L1 expression in patients with EGFR-Mutant non-small cell lung cancer. *Cancer Res. Treat.* **2018**, *50*, 95–102. [[CrossRef](#)]
- Christiani, D.C. Smoking and the molecular epidemiology of lung cancer. *Clin. Chest Med.* **2000**, *21*, 87–93. [[CrossRef](#)]
- Ha, S.Y.; Choi, S.-J.; Cho, J.H.; Choi, H.J.; Lee, J.; Jung, K.; Irwin, D.; Liu, X.; Lira, M.E.; Mao, M.; et al. Lung cancer in never-smoker Asian females is driven by oncogenic mutations, most often involving EGFR. *Oncotarget* **2015**, *6*, 5465–5474. [[CrossRef](#)] [[PubMed](#)]
- Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [[CrossRef](#)]
- Enstone, A.; Greaney, M.; Povsic, M.; Wyn, R.; Penrod, J.R.; Yuan, Y. The economic burden of small cell lung cancer: A systematic review of the literature. *Pharm. Open* **2017**, *2*, 139–152. [[CrossRef](#)]
- De Groot, P.M.; Wu, C.C.; Carter, B.W.; Munden, R.F. The epidemiology of lung cancer. *Transl. Lung Cancer Res.* **2018**, *7*, 220–233. [[CrossRef](#)]
- van der Meer, D.J.; Karim-Kos, H.E.; van der Mark, M.; Aben, K.K.H.; Bijlsma, R.M.; Rijneveld, A.W.; van der Graaf, W.T.A.; Husson, O. Incidence, survival, and mortality trends of cancers diagnosed in adolescents and young adults (15–39 Years): A population-based study in The Netherlands 1990–2016. *Cancers* **2020**, *18*, 3421. [[CrossRef](#)]
- Sacco, P.C.; Maione, P.; Guida, C.; Gridelli, C. The combination of new immunotherapy and radiotherapy: A new potential treatment for locally advanced non-small cell lung cancer. *Curr. Clin. Pharmacol.* **2017**, *12*, 4–10. [[CrossRef](#)]
- Hirsh, V. New developments in the treatment of advanced squamous cell lung cancer: Focus on afatinib. *OncoTargets Ther.* **2017**, *10*, 2513–2526. [[CrossRef](#)]
- Wang, H.; Zhang, J.; Shi, F.; Zhang, C.; Jiao, Q.; Zhu, H. Better cancer specific survival in young small cell lung cancer patients especially with AJCC stage III. *Oncotarget* **2017**, *8*, 34923–34934. [[CrossRef](#)]
- Arnold, B.N.; Thomas, D.C.; Rosen, J.E.; Salazar, M.C.; Blasberg, J.D.; Boffa, D.J.; Detterbeck, F.C.; Kim, A.W. Lung cancer in the very young: Treatment and survival in the national cancer data base. *J. Thorac. Oncol.* **2016**, *11*, 1121–1131. [[CrossRef](#)]
- Liu, M.; Cai, X.; Yu, W.; Lv, C.; Fu, X. Clinical significance of age at diagnosis among young non-small cell lung cancer patients under 40 years old: A population-based study. *Oncotarget* **2015**, *6*, 44963–44970. [[CrossRef](#)] [[PubMed](#)]
- Liu, B.; Quan, X.; Xu, C.; Lv, J.; Li, C.; Dong, L.; Liu, M. Lung cancer in young adults aged 35 years or younger: A full-scale analysis and review. *J. Cancer* **2019**, *10*, 3553–3559. [[CrossRef](#)] [[PubMed](#)]
- Rich, A.L.; Khakwani, A.; Free, C.M.; Tata, L.J.; Stanley, R.A.; Peake, M.D.; Hubbard, R.B.; Baldwin, D.R. Non-small cell lung cancer in young adults: Presentation and survival in the English National Lung Cancer Audit: QJM. *Int. J. Med.* **2015**, *108*, 891–897. [[CrossRef](#)] [[PubMed](#)]
- Arrieta, Ó.; Angulo, L.P.; Núñez-Valencia, C.; Dorantes-Gallareta, Y.; Macedo, E.O.; Martínez-López, D.; Alvarado, S.; Corona-Cruz, J.-F.; Oñate-Ocaña, L.F. Association of depression and anxiety on quality of life, treatment adherence, and prognosis in patients with advanced non-small cell lung cancer. *Ann. Surg. Oncol.* **2012**, *20*, 1941–1948. [[CrossRef](#)]
- Yan, X.; Chen, X.; Li, M.; Zhang, P. Prevalence and risk factors of anxiety and depression in Chinese patients with lung cancer: A cross-sectional study. *Cancer Manag. Res.* **2019**, *11*, 4347–4356. [[CrossRef](#)]
- Johnson, C.G.; Brodsky, J.L.; Cataldo, J.K. Lung cancer stigma, anxiety, depression, and quality of life. *J. Psychosoc. Oncol.* **2014**, *32*, 59–73. [[CrossRef](#)]
- Park, S.; Kang, C.H.; Hwang, Y.; Seong, Y.W.; Lee, H.J.; Park, I.K.; Kim, Y.T. Risk factors for postoperative anxiety and depression after surgical treatment for lung cancer. *Eur. J. Cardiothorac. Surg.* **2016**, *49*, e16–e21. [[CrossRef](#)]
- Ting, C.-T.; Kuo, C.-J.; Hu, H.-Y.; Lee, Y.-L.; Tsai, T.-H. Prescription frequency and patterns of Chinese herbal medicine for liver cancer patients in Taiwan: A cross-sectional analysis of the National Health Insurance Research Database. *BMC Complement. Altern. Med.* **2017**, *17*, 1–11. [[CrossRef](#)]
- Jung, J.Y.; Lee, J.M.; Kim, M.S.; Shim, Y.M.; Zo, J.I.; Yun, Y.H. Comparison of fatigue, depression, and anxiety as factors affecting posttreatment health-related quality of life in lung cancer survivors. *Psych. Oncol.* **2018**, *27*, 465–470. [[CrossRef](#)]

22. Ambrogi, F.; Biganzoli, E.; Boracchi, P. Multiple correspondence analysis in S-PLUS. *Comput. Methods Programs Biomed.* **2005**, *79*, 161–167. [[CrossRef](#)]
23. Shrivastav, M.; Iaizzo, P. Discrimination of ischemia and normal sinus rhythm for cardiac signals using a modified k means clustering algorithm. In Proceedings of the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 22–26 August 2007; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2007; Volume 2007, pp. 3856–3859.
24. Saatchi, M.; McClure, M.C.; McKay, S.D.; Rolf, M.M.; Kim, J.; Decker, J.E.; Taxis, T.M.; Chapple, R.H.; Ramey, H.R.; Northcutt, S.L.; et al. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genet. Sel. Evol. (GSE)* **2011**, *43*, 40. [[CrossRef](#)] [[PubMed](#)]
25. Wang, J. Consistent selection of the number of clusters via cross validation. *Biometrika* **2010**, *97*, 893–904. [[CrossRef](#)]
26. Hagen, K.B.; Aas, T.; Kvaloy, J.T.; Eriksen, H.R.; Soiland, H.; Lind, R. Fatigue, anxiety and depression overrule the role of oncological treatment in predicting self-reported health complaints in women with breast cancer compared to healthy controls. *Breast* **2016**, *28*, 100–106. [[CrossRef](#)] [[PubMed](#)]
27. Gorman, J.R.; Su, H.I.; Roberts, S.C.; Dominick, S.A.; Malcarne, V.L. Experiencing reproductive concerns as a female cancer survivor is associated with depression. *Cancer* **2015**, *121*, 935–942. [[CrossRef](#)] [[PubMed](#)]
28. King, G.; Zeng, L. Logistic Regression in Rare Events Data. *Politi. Anal.* **2001**, *9*, 137–163. [[CrossRef](#)]
29. Westphal, C. Logistic regression for extremely rare events: The case of school shootings. *SSRN Electron. J.* **2013**. [[CrossRef](#)]
30. Nations, J.A.; Nathan, S.D. Comorbidities of Advanced Lung Disease. *Mt. Sinai J. Med. A J. Transl. Pers. Med.* **2009**, *76*, 53–62. [[CrossRef](#)]
31. Sculier, J.P.; Botta, I.; Bucalau, A.M.; Compagnie, M.; Eskenazi, A.; Fischler, R.; Gorham, J.; Mans, L.; Rozen, L.; Speybrouck, S.; et al. Medical anticancer treatment of lung cancer associated with comorbidities: A review. *Lung Cancer* **2015**, *87*, 241–248. [[CrossRef](#)] [[PubMed](#)]
32. Paal, B.V. *A Comparison of Different Methods for Modelling Rare Events Data*; Universiteit Gent: Brussel, Belgium, 2014.
33. Seib, C.; Porter-Steele, J.; Ng, S.K.; Turner, J.; McGuire, A.; McDonald, N.; Balaam, S.; Yates, P.; McCarthy, A.; Anderson, D. Life stress and symptoms of anxiety and depression in women after cancer: The mediating effect of stress appraisal and coping. *Psychooncology* **2018**, *27*, 1787–1794. [[CrossRef](#)] [[PubMed](#)]
34. Hong-Jhe, C.; Chin-Yuan, K.; Ming-Shium, T.; Fu-Wei, W.; Ru-Yih, C.; Kuang-Chieh, H.; Hsiang-Ju, P.; Ming-Yueh, C.; Pan-Ming, C.; Chih-Chuan, P. The incidence and risk of osteoporosis in patients with anxiety disorder: A Population-based retrospective cohort study. *Medicine* **2016**, *95*, e4912. [[CrossRef](#)] [[PubMed](#)]
35. Yeh, M.J.; Chang, H.H. National health insurance in Taiwan. *Health Aff.* **2015**, *34*, 1067. [[CrossRef](#)]
36. Shi, Q.; Li, K.J.; Treuer, T.; Wang, B.C.M.; Gaich, C.L.; Lee, C.H.; Wu, W.S.; Furnback, W.; Tang, C.H. Estimating the response and economic burden of rheumatoid arthritis patients treated with biologic disease-modifying antirheumatic drugs in Taiwan using the National Health Insurance Research Database (NHIRD). *PLoS ONE* **2018**, *13*, e0193489. [[CrossRef](#)]
37. Lu, T.; Yang, X.; Huang, Y.; Zhao, M.; Li, M.; Ma, K.; Yin, J.; Zhan, C.; Wang, Q. Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades. *Cancer Manag. Res.* **2019**, *11*, 943–953. [[CrossRef](#)]